Asymptotically Free Sketching and Applications in Ridge Regression

Pratik Patil

MDS 2024

Thanks to collaborators

Daniel LeJeune

Hamid Javadi

Thanks to collaborators

Daniel LeJeune

Hamid Javadi

Rich Baraniuk

Ryan Tibshirani

Sketching

Sketching

- A *sketch* is a (random) linear projection that preserves geometry

Sketching

- A *sketch* is a (random) linear projection that preserves geometry

- Classical result (Johnson & Lindenstrauss, 1984): n points in $\mathbb{R}^p$ can be embedded by a linear map $\mathbf{S} \in \mathbb{R}^{p \times q}$ for $q \geq C\varepsilon^{-2} \log n$ such that

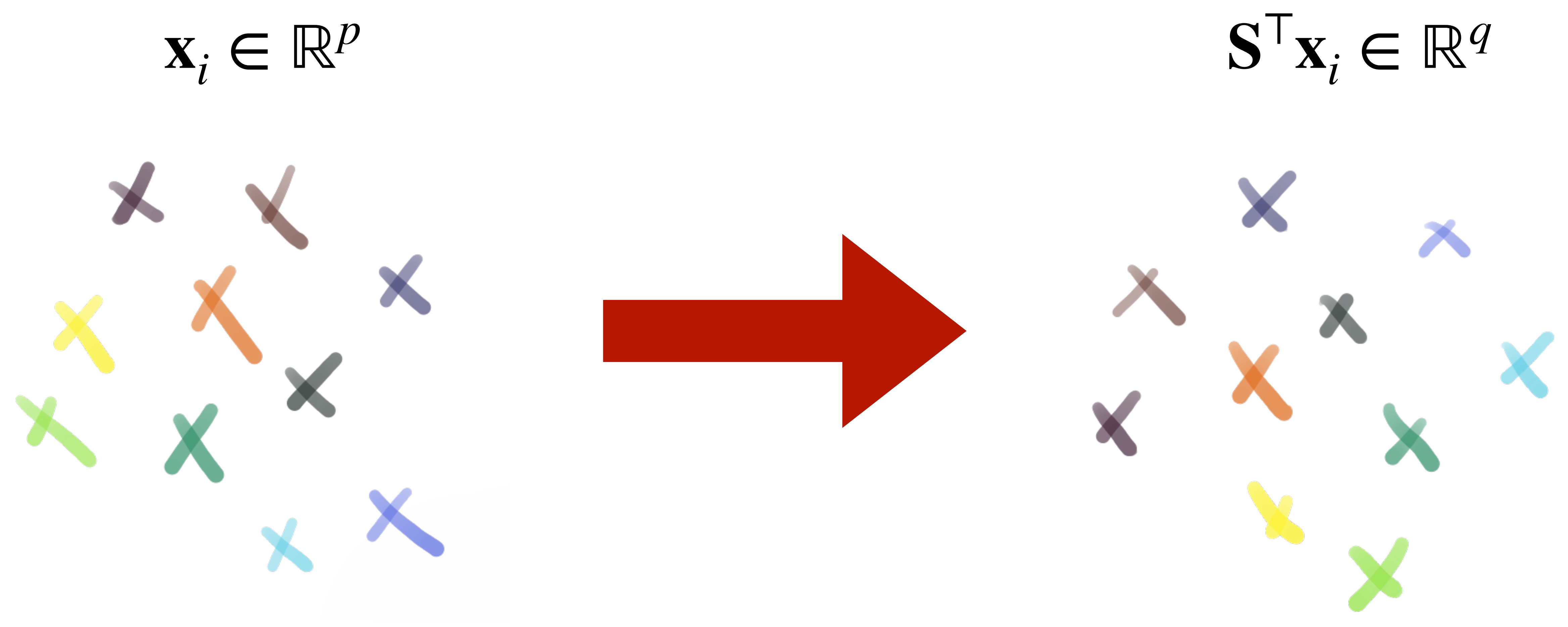
$$(1 - \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\mathbf{S}^\top \mathbf{x}_i - \mathbf{S}^\top \mathbf{x}_j\|_2 \leq (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Sketching

- A *sketch* is a (random) linear projection that preserves geometry

- Classical result (Johnson & Lindenstrauss, 1984): n points in $\mathbb{R}^p$ can be embedded by a linear map $\mathbf{S} \in \mathbb{R}^{p \times q}$ for $q \geq C\varepsilon^{-2} \log n$ such that

$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\mathbf{S}^\top \mathbf{x}_i - \mathbf{S}^\top \mathbf{x}_j\|_2 \leq (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2$$



Why sketch?

Why sketch?

- Example: Criteo 1TB dataset

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features:

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features: $p \sim 10^{11}$?

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features: $p \sim 10^{11}$?
    - More than 400GiB memory for regression coefficients alone

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features: $p \sim 10^{11}$?
    - More than 400GiB memory for regression coefficients alone
- Solution: the "hashing trick" **(sketching)**

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features: $p \sim 10^{11}$?
    - More than 400GiB memory for regression coefficients alone
- Solution: the "hashing trick" **(sketching)**
  - Take each 32-bit feature and hash it (e.g., 24-bit): 62e59341→73e059

Why sketch?

- Example: Criteo 1TB dataset
  - Binary ad click prediction, $n = 4{,}195{,}197{,}692$
  - 13 numerical features, 26 32-bit categorical features: $p \sim 10^{11}$?
    - More than 400GiB memory for regression coefficients alone
- Solution: the "hashing trick" **(sketching)**
  - Take each 32-bit feature and hash it (e.g., 24-bit): 62e59341→73e059
  - Use sum of one-hot encodings as features: $p \sim 10^7$, feasible

Where to find sketches

Where to find sketches

- **Computer science**

Where to find sketches

- **Computer science**
    - Find frequent items in data streams (Bloom filters, count(-min) sketch)

Where to find sketches

- **Computer science**
  - Find frequent items in data streams (Bloom filters, count(-min) sketch)
  - Fast database querying (locality sensitive hashing)

Where to find sketches

- **Computer science**
  - Find frequent items in data streams (Bloom filters, count(-min) sketch)
  - Fast database querying (locality sensitive hashing)
  - $\mathbf{S}$ built from hash functions, fast to apply and approximately invert

Where to find sketches

- **Computer science**
  - Find frequent items in data streams (Bloom filters, count(-min) sketch)
  - Fast database querying (locality sensitive hashing)
  - $\mathbf{S}$ built from hash functions, fast to apply and approximately invert
- **Compressed sensing**

Where to find sketches

- **Computer science**
  - Find frequent items in data streams (Bloom filters, count(-min) sketch)
  - Fast database querying (locality sensitive hashing)
  - $\mathbf{S}$ built from hash functions, fast to apply and approximately invert
- **Compressed sensing**
  - Store sparse high-dimensional measurements to recover downstream

Where to find sketches

- **Computer science**
  - Find frequent items in data streams (Bloom filters, count(-min) sketch)
  - Fast database querying (locality sensitive hashing)
  - $\mathbf{S}$ built from hash functions, fast to apply and approximately invert
- **Compressed sensing**
  - Store sparse high-dimensional measurements to recover downstream
  - $\mathbf{S}$ chosen to maximize recovery quality

Where to find sketches

- **Numerical linear algebra & optimization**

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently
  - S chosen to minimize computational and memory costs

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently
  - S chosen to minimize computational and memory costs
- **Statistics & machine learning**

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently
  - $\mathbf{S}$ chosen to minimize computational and memory costs
- **Statistics & machine learning**
  - $\mathbf{y} = \mathbf{X}\mathbf{b}$: regression labels are a sketch of underlying coefficients

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently
  - $\mathbf{S}$ chosen to minimize computational and memory costs
- **Statistics & machine learning**
  - $\mathbf{y} = \mathbf{X}\mathbf{b}$: regression labels are a sketch of underlying coefficients
  - $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$: data is a sketch of population covariance

Where to find sketches

- **Numerical linear algebra & optimization**
  - Approximate the solution to a (sub-)problem efficiently
  - $\mathbf{S}$ chosen to minimize computational and memory costs
- **Statistics & machine learning**
  - $\mathbf{y} = \mathbf{X}\mathbf{b}$: regression labels are a sketch of underlying coefficients
  - $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$: data is a sketch of population covariance
  - $\mathbf{S} = \mathbf{X}^\top$ or $\mathbf{S} = \mathbf{Z}^\top$ is determined by nature and may not be observed

What do sketches look like?

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top\mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
    - Exemplar: uniformly random matrix with orthonormal columns

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
  - Exemplar: uniformly random matrix with orthonormal columns
  - Faster: subsampled randomized Fourier transforms (e.g., SRHT)

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
  - Exemplar: uniformly random matrix with orthonormal columns
  - Faster: subsampled randomized Fourier transforms (e.g., SRHT)
  - Fastest: subsampling matrix

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
    - Exemplar: uniformly random matrix with orthonormal columns
    - Faster: subsampled randomized Fourier transforms (e.g., SRHT)
    - Fastest: subsampling matrix
- **Independently random sketches**

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
    - Exemplar: uniformly random matrix with orthonormal columns
    - Faster: subsampled randomized Fourier transforms (e.g., SRHT)
    - Fastest: subsampling matrix
- **Independently random sketches**
    - Exemplar: i.i.d. (sub-)Gaussian matrix

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
  - Exemplar: uniformly random matrix with orthonormal columns
  - Faster: subsampled randomized Fourier transforms (e.g., SRHT)
  - Fastest: subsampling matrix
- **Independently random sketches**
  - Exemplar: i.i.d. (sub-)Gaussian matrix
  - Faster: sum of independent hashing functions (e.g., CountSketch)

What do sketches look like?

- **Orthonormal sketches** $\mathbf{S}^\top \mathbf{S} \propto \mathbf{I}_q$ $(q \times q)$
  - Exemplar: uniformly random matrix with orthonormal columns ✓
  - Faster: subsampled randomized Fourier transforms (e.g., SRHT) ✓
  - Fastest: subsampling matrix ✗ requires additional assumptions
- **Independently random sketches**
  - Exemplar: i.i.d. (sub-)Gaussian matrix ✓
  - Faster: sum of independent hashing functions (e.g., CountSketch) ✓

Q: How does sketching affect the result in machine learning?

Classical sketching theory

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

- **Example:** ridge regression, $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \sigma \mathbf{z}$

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

- **Example:** ridge regression, $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \sigma\mathbf{z}$

$$\widehat{\mathbf{b}} = \mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}, \quad \widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}\cdot\mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}$$

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

- **Example:** ridge regression, $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \sigma\mathbf{z}$

$$\widehat{\mathbf{b}} = \text{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}, \quad \widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \text{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}$$

Fixed design excess risk: $R(\mathbf{b}) = \mathbb{E}_{\mathbf{z}}[\|\mathbf{X}(\mathbf{b} - \mathbf{b}^*)\|_2^2]$

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

- **Example:** ridge regression, $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \sigma\mathbf{z}$

$$\widehat{\mathbf{b}} = \mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}, \quad \widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}\cdot \mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}$$

Fixed design excess risk: $R(\mathbf{b}) = \mathbb{E}_{\mathbf{z}}[\|\mathbf{X}(\mathbf{b} - \mathbf{b}^*)\|_2^2]$

Theorem (Lu et al. 2013). Let $r = \mathrm{rank}(\mathbf{X})$ and $\mathbf{S}$ be SRHT. With high probability,

$$R(\widehat{\mathbf{b}}_{\mathbf{S}}) - R(\widehat{\mathbf{b}}) \le C\frac{r\log r}{q}R(\widehat{\mathbf{b}}).$$

Classical sketching theory

- **Most results:** if sketch size is larger than inherent dimensionality, then using a sketch results in negligible error in recovering the solution

- **Example:** ridge regression, $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \sigma\mathbf{z}$

$$\hat{\mathbf{b}} = \mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}, \quad \hat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}\cdot\mathrm{argmin}_{\mathbf{b}}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\right\}$$

Fixed design excess risk: $R(\mathbf{b}) = \mathbb{E}_{\mathbf{z}}[\|\mathbf{X}(\mathbf{b} - \mathbf{b}^*)\|_2^2]$

Theorem (Lu et al. 2013). Let $r = \mathrm{rank}(\mathbf{X})$ and $\mathbf{S}$ be SRHT. With high probability,

$$R(\hat{\mathbf{b}}_{\mathbf{S}}) - R(\hat{\mathbf{b}}) \leq C\frac{r\log r}{q}R(\hat{\mathbf{b}}).$$

What about small sketches?

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p - q}{p},$$

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

Bias: $$R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}.$$

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}.$$

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \boxed{\sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \boxed{\frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}}$$

equal when $\dfrac{1}{1+\lambda} = \dfrac{q}{p}$

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \boxed{\sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \boxed{\frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}}$$

equal when $\frac{1}{1+\lambda} = \frac{q}{p}$

Does $\mathbb{E}[\widehat{\mathbf{b}}_{\mathbf{S}}] = \widehat{\mathbf{b}}$?

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$ $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

1

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}$$

equal when $\frac{1}{1+\lambda} = \frac{q}{p}$

Does $\mathbb{E}[\widehat{\mathbf{b}}_{\mathbf{S}}] = \widehat{\mathbf{b}}$?

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$ (1), $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$ (2), and $q \leq p$. Then for $\lambda = 0$,

$$\mathbb{E}_{\mathbf{S}}[R(\widehat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_{\mathbf{S}}[\widehat{\mathbf{b}}_{\mathbf{S}}]) = \boxed{\sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \boxed{\frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}}$$

equal when $\frac{1}{1+\lambda} = \frac{q}{p}$

Does $\mathbb{E}[\widehat{\mathbf{b}}_{\mathbf{S}}] = \widehat{\mathbf{b}}$?

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

1 2 3

$$\mathbb{E}_\mathbf{S}[R(\widehat{\mathbf{b}}_\mathbf{S})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

$$\text{Bias: } R(\mathbb{E}_\mathbf{S}[\widehat{\mathbf{b}}_\mathbf{S}]) = \sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}.$$

- Compare to ridge regression:

$$R(\widehat{\mathbf{b}}) = \frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}$$

equal when $\frac{1}{1+\lambda} = \frac{q}{p}$

Does $\mathbb{E}[\widehat{\mathbf{b}}_\mathbf{S}] = \widehat{\mathbf{b}}$?

What about small sketches?

- **Theorem** (Thanei et al. 2017). Let $\mathbf{S}$ be i.i.d. $\mathcal{N}(0, q^{-1})$, $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, and $q \leq p$. Then for $\lambda = 0$,

1 2 3 4

$$\mathbb{E}_{\mathbf{S}}[R(\hat{\mathbf{b}}_{\mathbf{S}})] = \sigma^2 q + \|\mathbf{b}^*\|_2^2 \frac{p-q}{p},$$

Bias: $R(\mathbb{E}_{\mathbf{S}}[\hat{\mathbf{b}}_{\mathbf{S}}]) = \sigma^2 \frac{q^2}{p} + \|\mathbf{b}^*\|_2^2 \frac{(p-q)^2}{p^2}.$

- Compare to ridge regression:

$$R(\hat{\mathbf{b}}) = \frac{\sigma^2 p}{(1+\lambda)^2} + \|\mathbf{b}^*\|_2^2 \frac{\lambda^2}{(1+\lambda)^2}$$

equal when $\frac{1}{1+\lambda} = \frac{q}{p}$

Does $\mathbb{E}[\hat{\mathbf{b}}_{\mathbf{S}}] = \hat{\mathbf{b}}$?

$\mathbb{E}$(sketching) = ridge?

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$
  - $\widehat{\mathbf{b}} = \frac{1}{1+\lambda}(\mathbf{b}^* + \sigma \mathbf{z})$ while $\widehat{\mathbf{b}}_\mathbf{S} = \mathbf{S}(\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top (\mathbf{b}^* + \sigma \mathbf{z})$

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$
  - $\widehat{\mathbf{b}} = \frac{1}{1+\lambda}(\mathbf{b}^* + \sigma\mathbf{z})$ while $\widehat{\mathbf{b}}_\mathbf{S} = \mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top(\mathbf{b}^* + \sigma\mathbf{z})$
  - By rotational symmetry, $\mathbb{E}[\mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top] = c_\lambda\mathbf{I}_p$, $c_\lambda < 1$

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$
  - $\widehat{\mathbf{b}} = \frac{1}{1+\lambda}(\mathbf{b}^* + \sigma\mathbf{z})$ while $\widehat{\mathbf{b}}_{\mathbf{S}} = \boxed{\mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top}(\mathbf{b}^* + \sigma\mathbf{z})$
  - By rotational symmetry, $\mathbb{E}[\mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top] = c_\lambda\mathbf{I}_p$, $c_\lambda < 1$
  - So for any $\widehat{\mathbf{b}}_{\mathbf{S}}$ at $\lambda > 0$, $\mathbb{E}[\widehat{\mathbf{b}}_{\mathbf{S}}] = \widehat{\mathbf{b}}$ for $\widehat{\mathbf{b}}$ at some $\lambda' > 0$ ✓

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$
  - $\widehat{\mathbf{b}} = \frac{1}{1+\lambda}(\mathbf{b}^* + \sigma\mathbf{z})$ while $\widehat{\mathbf{b}}_\mathbf{S} = \boxed{\mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top}(\mathbf{b}^* + \sigma\mathbf{z})$
  - By rotational symmetry, $\mathbb{E}[\mathbf{S}(\mathbf{S}^\top\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top] = c_\lambda\mathbf{I}_p$, $c_\lambda < 1$
  - So for any $\widehat{\mathbf{b}}_\mathbf{S}$ at $\lambda > 0$, $\mathbb{E}[\widehat{\mathbf{b}}_\mathbf{S}] = \widehat{\mathbf{b}}$ for $\widehat{\mathbf{b}}$ at some $\lambda' > 0$ ✓
- What about arbitrary $\mathbf{X}$?

$\mathbb{E}$(sketching) = ridge?

- Orthogonal design setting $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$, i.i.d. Gaussian $\mathbf{S}$
  - $\widehat{\mathbf{b}} = \frac{1}{1+\lambda}(\mathbf{b}^* + \sigma \mathbf{z})$ while $\widehat{\mathbf{b}}_\mathbf{S} = \mathbf{S}(\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top (\mathbf{b}^* + \sigma \mathbf{z})$
  - By rotational symmetry, $\mathbb{E}[\mathbf{S}(\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top] = c_\lambda \mathbf{I}_p$, $c_\lambda < 1$
  - So for any $\widehat{\mathbf{b}}_\mathbf{S}$ at $\lambda > 0$, $\mathbb{E}[\widehat{\mathbf{b}}_\mathbf{S}] = \widehat{\mathbf{b}}$ for $\widehat{\mathbf{b}}$ at some $\lambda' > 0$ ✓
- What about arbitrary $\mathbf{X}$?
  - We need a different "expectation"

Asymptotic equivalences

Asymptotic equivalences

- Also known as "deterministic equivalences" in random matrix theory

Asymptotic equivalences

- Also known as "deterministic equivalences" in random matrix theory

- **Definition.** Two sequences of matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ are *asymptotically equivalent,* written $\mathbf{A}_n \simeq \mathbf{B}_n$, if for any sequence $\mathbf{\Theta}_n$ independent of $\mathbf{A}_n$ and $\mathbf{B}_n$ with bounded trace norm,

$$\lim_{n\to\infty} \mathrm{tr}[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)] = 0 \quad \text{almost surely.}$$

Asymptotic equivalences

- Also known as "deterministic equivalences" in random matrix theory

- **Definition.** Two sequences of matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ are *asymptotically equivalent,* written $\mathbf{A}_n \simeq \mathbf{B}_n$, if for any sequence $\mathbf{\Theta}_n$ independent of $\mathbf{A}_n$ and $\mathbf{B}_n$ with bounded trace norm,

$$\lim_{n\to\infty} \mathrm{tr}[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)] = 0 \quad \text{almost surely.}$$

- Typical usage: $\mathbf{A}_n$ is complicated and $\mathbf{B}_n$ is simple, but $\mathbf{A}_n \simeq \mathbf{B}_n$, so we can use $\mathbf{B}_n$ to understand $\mathbf{A}_n$

Asymptotic equivalences

- Also known as "deterministic equivalences" in random matrix theory

- **Definition.** Two sequences of matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ are *asymptotically equivalent,* written $\mathbf{A}_n \simeq \mathbf{B}_n$, if for any sequence $\mathbf{\Theta}_n$ independent of $\mathbf{A}_n$ and $\mathbf{B}_n$ with bounded trace norm,

$$\lim_{n\to\infty} \mathrm{tr}[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)] = 0 \quad \text{almost surely.}$$

- Typical usage: $\mathbf{A}_n$ is complicated and $\mathbf{B}_n$ is simple, but $\mathbf{A}_n \simeq \mathbf{B}_n$, so we can use $\mathbf{B}_n$ to understand $\mathbf{A}_n$

- $\mathbf{A}_n \simeq \mathbf{B}_n$ is analogous to $\mathbb{E}[\mathbf{A}] = \mathbb{E}[\mathbf{B}]$, but single-instance

Calculus of asymptotic equivalences

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Product: $\mathbf{A}_n \simeq \mathbf{B}_n, (\mathbf{A}_n, \mathbf{B}_n) \perp (\mathbf{C}_n, \mathbf{D}_n) \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Product: $\mathbf{A}_n \simeq \mathbf{B}_n, (\mathbf{A}_n, \mathbf{B}_n) \perp (\mathbf{C}_n, \mathbf{D}_n) \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$
  - Elements: $\mathbf{A}_n \simeq \mathbf{B}_n \implies [\mathbf{A}_n]_{ij} - [\mathbf{B}_n]_{ij} \xrightarrow{\text{a.s.}} 0$

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Product: $\mathbf{A}_n \simeq \mathbf{B}_n, (\mathbf{A}_n, \mathbf{B}_n) \perp (\mathbf{C}_n, \mathbf{D}_n) \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$
  - Elements: $\mathbf{A}_n \simeq \mathbf{B}_n \implies [\mathbf{A}_n]_{ij} - [\mathbf{B}_n]_{ij} \xrightarrow{\text{a.s.}} 0$
  - Derivative: $f(\mathbf{A}_n; z) \simeq g(\mathbf{B}_n; z) \implies f'(\mathbf{A}_n; z) \simeq g'(\mathbf{B}_n; z)$

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Product: $\mathbf{A}_n \simeq \mathbf{B}_n, (\mathbf{A}_n, \mathbf{B}_n) \perp (\mathbf{C}_n, \mathbf{D}_n) \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$
  - Elements: $\mathbf{A}_n \simeq \mathbf{B}_n \implies [\mathbf{A}_n]_{ij} - [\mathbf{B}_n]_{ij} \xrightarrow{\text{a.s.}} 0$
  - Derivative: $f(\mathbf{A}_n; z) \simeq g(\mathbf{B}_n; z) \implies f'(\mathbf{A}_n; z) \simeq g'(\mathbf{B}_n; z)$
- **Not nonlinear ops:** $\mathbf{A}_n \simeq \mathbf{B}_n \not\Longrightarrow \mathbf{A}_n^k \simeq \mathbf{B}_n^k$

Calculus of asymptotic equivalences

- Asymptotic equivalences admit a calculus (Dobriban and Sheng, 2021)
  - Sum: $\mathbf{A}_n \simeq \mathbf{B}_n, \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Product: $\mathbf{A}_n \simeq \mathbf{B}_n, (\mathbf{A}_n, \mathbf{B}_n) \perp (\mathbf{C}_n, \mathbf{D}_n) \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$
  - Elements: $\mathbf{A}_n \simeq \mathbf{B}_n \implies [\mathbf{A}_n]_{ij} - [\mathbf{B}_n]_{ij} \xrightarrow{\text{a.s.}} 0$
  - Derivative: $f(\mathbf{A}_n; z) \simeq g(\mathbf{B}_n; z) \implies f'(\mathbf{A}_n; z) \simeq g'(\mathbf{B}_n; z)$
- **Not nonlinear ops:** $\mathbf{A}_n \simeq \mathbf{B}_n \not\Longrightarrow \mathbf{A}_n^k \simeq \mathbf{B}_n^k$
  - Analogous: $\mathbb{E}[\mathbf{A}] = \mathbb{E}[\mathbf{B}] \not\Longrightarrow \mathbb{E}[\mathbf{A}^k] = \mathbb{E}[\mathbf{B}^k]$

The sketched pseudoinverse

The sketched pseudoinverse

- Sketched ridge: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \mathrm{argmin}_{\mathbf{b}} \{ \|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \}$

The sketched pseudoinverse

- Sketched ridge: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \operatorname{argmin}_{\mathbf{b}} \{ \|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \}$
  - Closed form: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top \mathbf{X}^\top \mathbf{y}$

The sketched pseudoinverse

- Sketched ridge: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \operatorname{argmin}_{\mathbf{b}} \{ \|\mathbf{y} - \mathbf{XSb}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \}$
  - Closed form: $\widehat{\mathbf{b}}_{\mathbf{S}} = \boxed{\mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top} \mathbf{X}^\top \mathbf{y}$

The sketched pseudoinverse

- Sketched ridge: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \operatorname{argmin}_{\mathbf{b}} \{ \|\mathbf{y} - \mathbf{XSb}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \}$
    - Closed form: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top \mathbf{X}^\top \mathbf{y}$
- **Definition.** Given a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ and sketching matrix $\mathbf{S} \in \mathbb{R}^{p \times q}$, its *sketched pseudoinverse* with regularization λ is

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top.$$

The sketched pseudoinverse

- Sketched ridge: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S} \cdot \operatorname{argmin}_{\mathbf{b}} \{ \|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \}$
    - Closed form: $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \mathbf{X}^\top \mathbf{y}$
- **Definition.** Given a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ and sketching matrix $\mathbf{S} \in \mathbb{R}^{p \times q}$, its *sketched pseudoinverse* with regularization λ is

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top.$$

    - Called "pseudoinverse" because when $\mathbf{S}$ has orthonormal columns and $\lambda \to 0$, it is the Moore–Penrose pseudoinverse of $\mathbf{S}\mathbf{S}^\top \mathbf{A}\mathbf{S}\mathbf{S}^\top$

A first-order asymptotic equivalence

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p \to \infty} \alpha \in (0, \infty)$

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p \to \infty} \alpha \in (0, \infty)$

- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0$, $\mathbb{E}[S_{ij}^2] = \frac{1}{q}$, $\mathbb{E}[|\sqrt{q} S_{ij}|^{8+\delta}] < \infty$

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p \to \infty} \alpha \in (0, \infty)$
- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0$, $\mathbb{E}[S_{ij}^2] = \frac{1}{q}$, $\mathbb{E}[\,|\sqrt{q} S_{ij}|^{8+\delta}] < \infty$
- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A} \mathbf{S})$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A} \mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

where μ is the most positive solution to

$$\lambda = \mu\left(1 - \frac{1}{q} \operatorname{tr}[\mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]\right).$$

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p\to\infty} \alpha \in (0,\infty)$
- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0$, $\mathbb{E}[S_{ij}^2] = \frac{1}{q}$, $\mathbb{E}[\,|\sqrt{q}S_{ij}|^{8+\delta}] < \infty$
- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

 where μ is the most positive solution to

$$\lambda = \mu\Big(1 - \frac{1}{q}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]\Big).$$

- **Proof idea:** classical RMT techniques (generalized Marchenko–Pastur), analytic continuation

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p\to\infty} \alpha \in (0,\infty)$

- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0,\ \mathbb{E}[S_{ij}^2] = \frac{1}{q},\ \mathbb{E}[|\sqrt{q}S_{ij}|^{8+\delta}] < \infty$ 1*

- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

 where μ is the most positive solution to

$$\lambda = \mu\left(1 - \frac{1}{q}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]\right).$$

- **Proof idea:** classical RMT techniques (generalized Marchenko–Pastur), analytic continuation

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p\to\infty} \alpha \in (0,\infty)$

- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0$, $\mathbb{E}[S_{ij}^2] = \frac{1}{q}$, $\mathbb{E}[|\sqrt{q}S_{ij}|^{8+\delta}] < \infty$ 1*

- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm 2 independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

  where μ is the most positive solution to

$$\lambda = \mu\Big(1 - \frac{1}{q}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]\Big).$$

- **Proof idea:** classical RMT techniques (generalized Marchenko–Pastur), analytic continuation

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p\to\infty} \alpha \in (0,\infty)$ 3
- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0$, $\mathbb{E}[S_{ij}^2] = \frac{1}{q}$, $\mathbb{E}[|\sqrt{q}S_{ij}|^{8+\delta}] < \infty$ 1*
- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm 2 independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$,

$$\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

where μ is the most positive solution to

$$\lambda = \mu\left(1 - \frac{1}{q}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]\right).$$

- **Proof idea:** classical RMT techniques (generalized Marchenko–Pastur), analytic continuation

A first-order asymptotic equivalence

- Proportional asymptotics, $\alpha = \frac{q}{p}$: $\lim_{p\to\infty} \alpha \in (0,\infty)$ 3
- I.i.d. sketching: $\mathbb{E}[S_{ij}] = 0,\ \mathbb{E}[S_{ij}^2] = \frac{1}{q},\ \mathbb{E}[|\sqrt{q}S_{ij}|^{8+\delta}] < \infty$ 1*
- **Theorem** (LeJeune, **PP**, et al., 2024). For $\mathbf{A}$ with uniformly bounded in operator norm 2 independent of $\mathbf{S}$ and any $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$, 4

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

where μ is the most positive solution to

$$\lambda = \mu\left(1 - \frac{1}{q}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]\right).$$

- **Proof idea:** classical RMT techniques (generalized Marchenko–Pastur), analytic continuation

What about other sketches?

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
  - Does the same equivalence hold?

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
  - Does the same equivalence hold?

- **Theorem** (LeJeune, **PP**, et al., 2024). Let $\mathbf{S}\mathbf{S}^\top$ be almost surely asymptotically free from $\mathbf{A}$ and $\mathbf{\Theta}$ and have analytic S-transform $S_{\mathbf{S}\mathbf{S}^\top}$. Then for $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{S})$

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

  where μ solves

$$\mu = \lambda S_{\mathbf{S}\mathbf{S}^\top}(-\frac{1}{p}\operatorname{tr}[\mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]).$$

- **Proof idea:** fusion of free probability theory + Jacobi's formula + differential calculus

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
  - Does the same equivalence hold?

1

- **Theorem** (LeJeune, **PP**, et al., 2024). Let $\mathbf{S}\mathbf{S}^\top$ be almost surely asymptotically free from $\mathbf{A}$ and $\boldsymbol{\Theta}$ and have analytic S-transform $S_{\mathbf{S}\mathbf{S}^\top}$. Then for $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top\mathbf{A}\mathbf{S})$

$$\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

  where μ solves

$$\mu = \lambda S_{\mathbf{S}\mathbf{S}^\top}(-\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]).$$

- **Proof idea:** fusion of free probability theory + Jacobi's formula + differential calculus

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
  - Does the same equivalence hold?

1

- **Theorem** (LeJeune, **PP**, et al., 2024). Let $\mathbf{S}\mathbf{S}^\top$ be almost surely asymptotically free from $\mathbf{A}$ and $\boldsymbol{\Theta}$ and have analytic S-transform $S_{\mathbf{S}\mathbf{S}^\top}$. Then for $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top\mathbf{A}\mathbf{S})$

2

$$\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

where μ solves

$$\mu = \lambda S_{\mathbf{S}\mathbf{S}^\top}(-\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]).$$

- **Proof idea:** fusion of free probability theory + Jacobi's formula + differential calculus

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
    - Does the same equivalence hold?

1

- **Theorem** (LeJeune, **PP**, et al., 2024). Let $\mathbf{S}\mathbf{S}^\top$ be almost surely asymptotically free from $\mathbf{A}$ and $\mathbf{\Theta}$ and have analytic S-transform $S_{\mathbf{S}\mathbf{S}^\top}$. Then for $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top\mathbf{A}\mathbf{S})$

2

$$\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

where μ solves

α is implicit

3

$$\mu = \lambda S_{\mathbf{S}\mathbf{S}^\top}\left(-\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]\right).$$

- **Proof idea:** fusion of free probability theory + Jacobi's formula + differential calculus

What about other sketches?

- Orthogonal, SRHT, CountSketch, etc. are not i.i.d. sketches
  - Does the same equivalence hold?

1

- **Theorem** (LeJeune, **PP**, et al., 2024). Let $\mathbf{S}\mathbf{S}^\top$ be almost surely asymptotically free from $\mathbf{A}$ and $\mathbf{\Theta}$ and have analytic S-transform $S_{\mathbf{S}\mathbf{S}^\top}$. Then for $\lambda > -\lim \lambda_{\min}(\mathbf{S}^\top\mathbf{A}\mathbf{S})$ 4

2

$$\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

where μ solves α is implicit

3

$$\mu = \lambda S_{\mathbf{S}\mathbf{S}^\top}\left(-\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}]\right).$$

- **Proof idea:** fusion of free probability theory + Jacobi's formula + differential calculus

Orthogonal sketches

Orthogonal sketches

- **Corollary** (LeJeune, **PP**, et al., 2024). Let $\sqrt{\frac{p}{q}}\mathbf{S}$ have orthonormal columns. Then

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \gamma \mathbf{I}_p)^{-1},$$

where γ is the most positive solution to

$$\frac{1}{p}\mathrm{tr}[(\mathbf{A} + \mu \mathbf{I}_p)^{-1}](\gamma - \alpha\lambda) = 1 - \alpha.$$

Orthogonal sketches

- **Corollary** (LeJeune, **PP**, et al., 2024). Let $\sqrt{\frac{p}{q}}\mathbf{S}$ have orthonormal columns. Then

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \gamma \mathbf{I}_p)^{-1},$$

where γ is the most positive solution to

$$\frac{1}{p}\mathrm{tr}[(\mathbf{A} + \mu \mathbf{I}_p)^{-1}](\gamma - \alpha\lambda) = 1 - \alpha.$$

- **Less distortion:** $\lambda < \gamma < \mu$ for i.i.d. sketch when $\mu > 0$

Empirical concentration

Empirical concentration

- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$

Empirical concentration

- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$
- $\alpha = 0.8, \lambda = 1, \mu \approx 1.63, \gamma \approx 1.17$

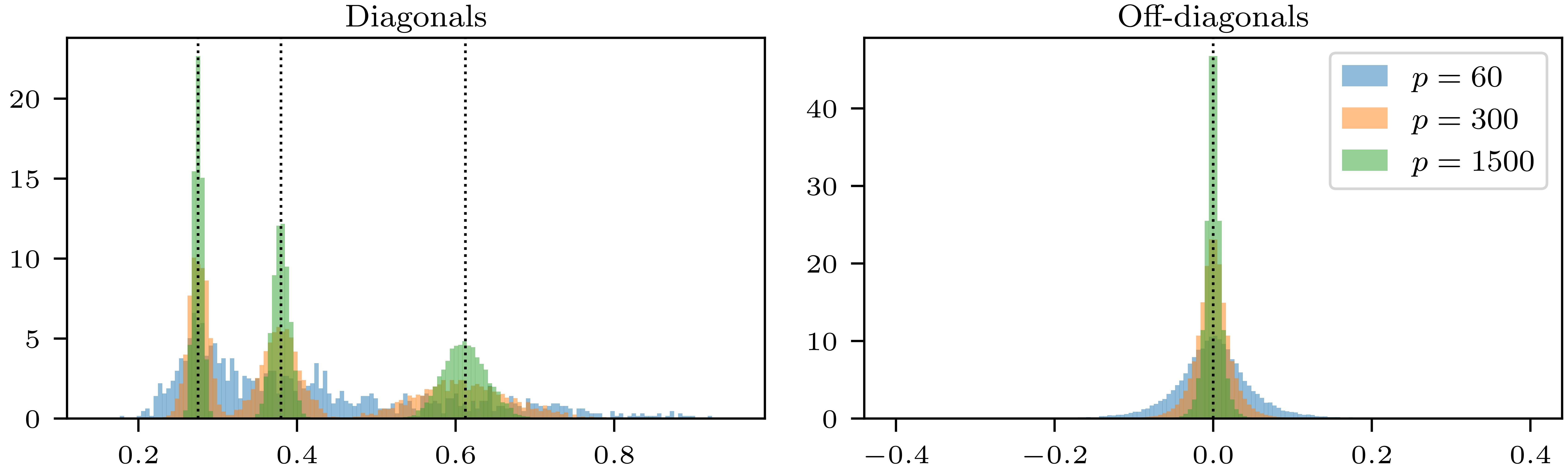
Empirical concentration

- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$
- $\alpha = 0.8$, $\lambda = 1$, $\mu \approx 1.63$, $\gamma \approx 1.17$
- Examine $[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top]_{ij} \simeq [(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]_{ij}$

Empirical concentration

- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$
- $\alpha = 0.8, \lambda = 1, \mu \approx 1.63, \gamma \approx 1.17$
- Examine $[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top]_{ij} \simeq [(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]_{ij}$

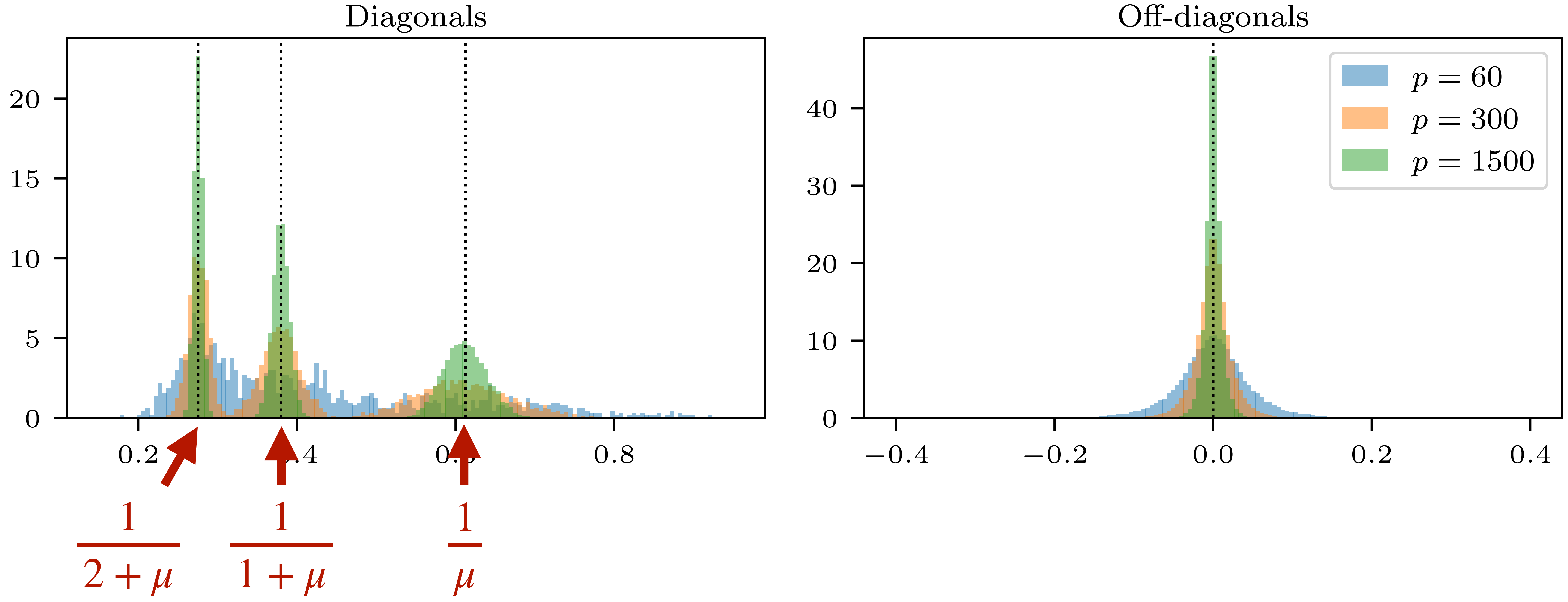
I.i.d. sketch



Empirical concentration

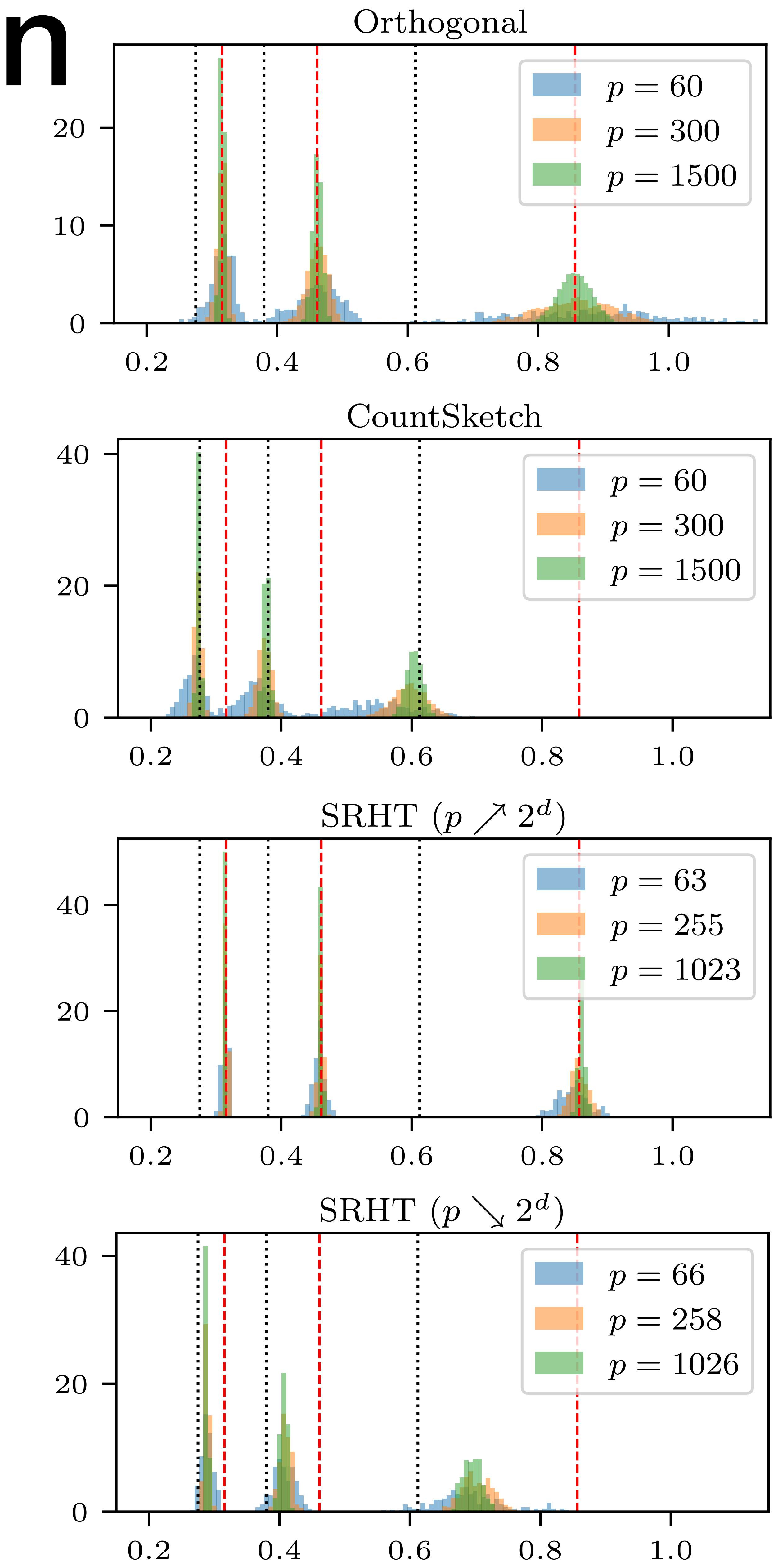
- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$
- $\alpha = 0.8, \lambda = 1, \mu \approx 1.63, \gamma \approx 1.17$
- Examine $[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top]_{ij} \simeq [(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]_{ij}$

I.i.d. sketch

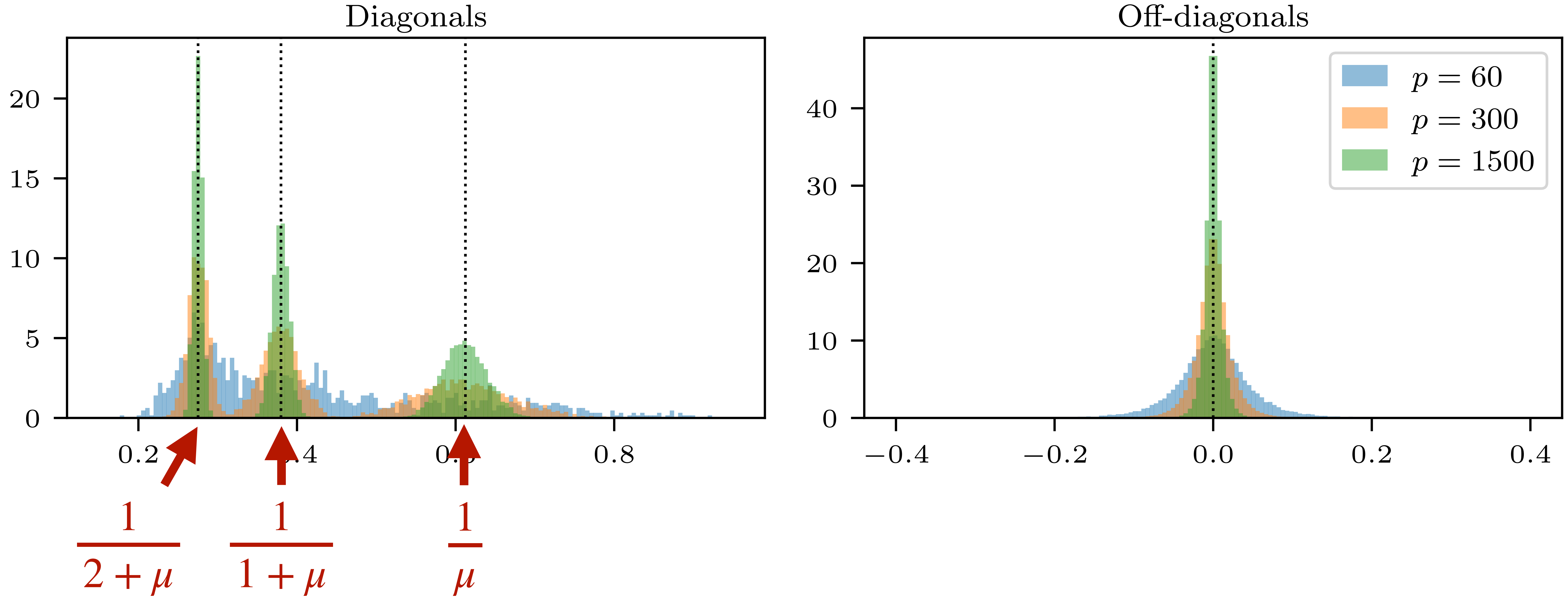


Empirical concentration

- Example: $\mathbf{A} = \mathrm{diag}(0,\ldots,1,\ldots,2,\ldots)$
- $\alpha = 0.8, \lambda = 1, \mu \approx 1.63, \gamma \approx 1.17$
- Examine $[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top]_{ij} \simeq [(\mathbf{A} + \mu \mathbf{I}_p)^{-1}]_{ij}$



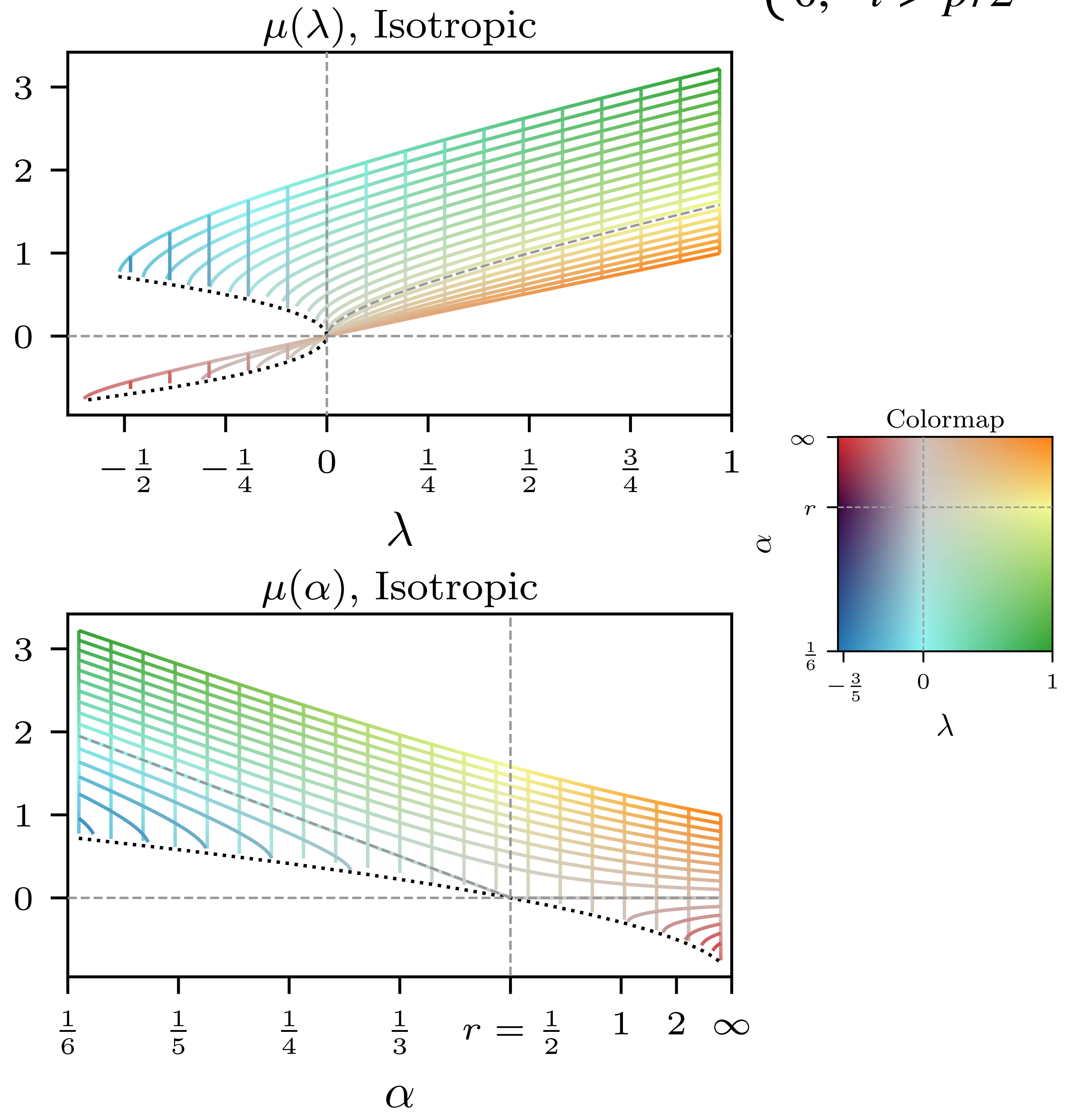
I.i.d. sketch



Some intuition about $\lambda \mapsto \mu$

Some intuition about $\lambda \mapsto \mu$

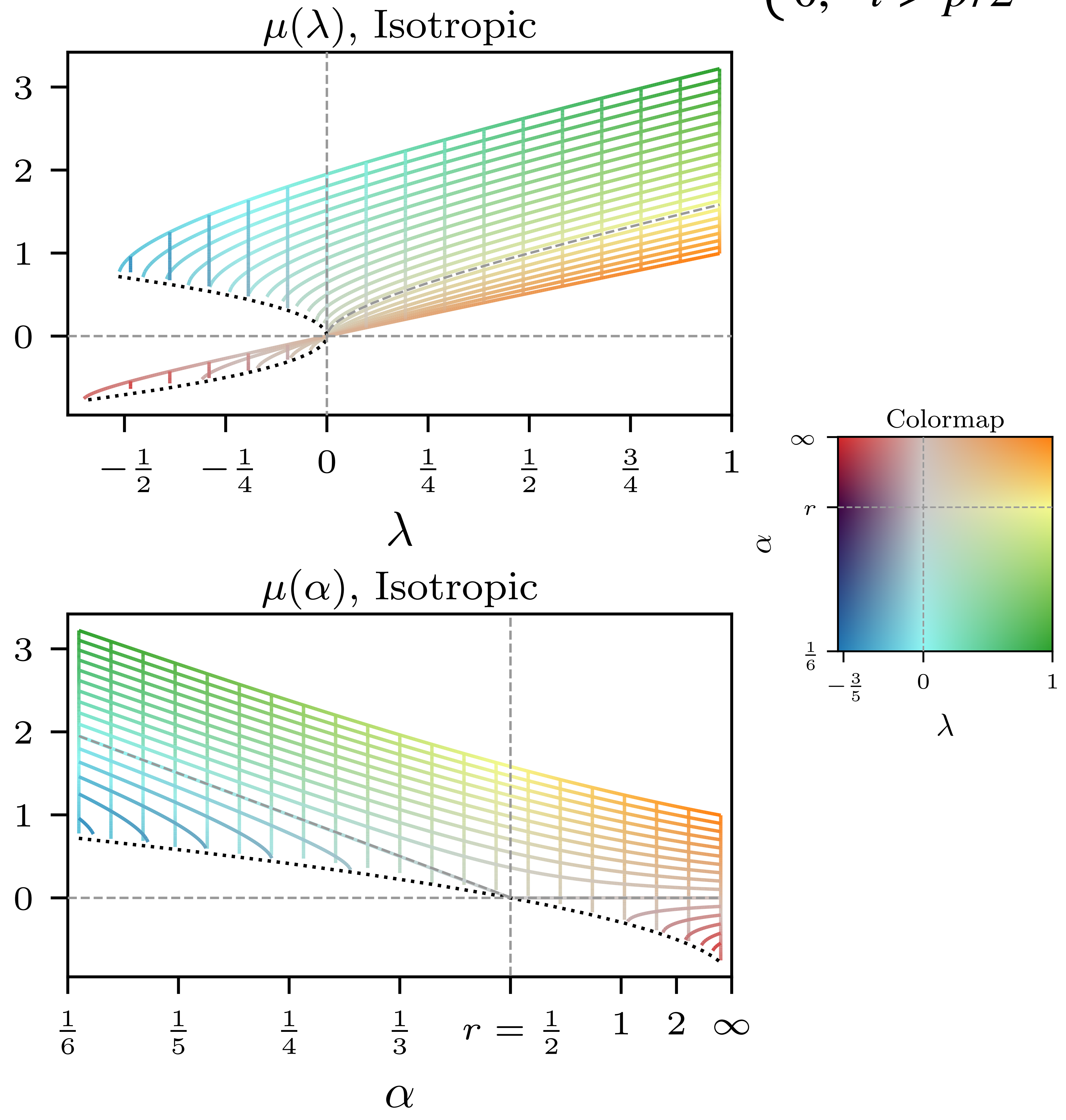
$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \le p/2 \\ 0, & i > p/2 \end{cases}$$



Some intuition about $\lambda \mapsto \mu$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \le p/2 \\ 0, & i > p/2 \end{cases}$$

- Concave and increasing in λ

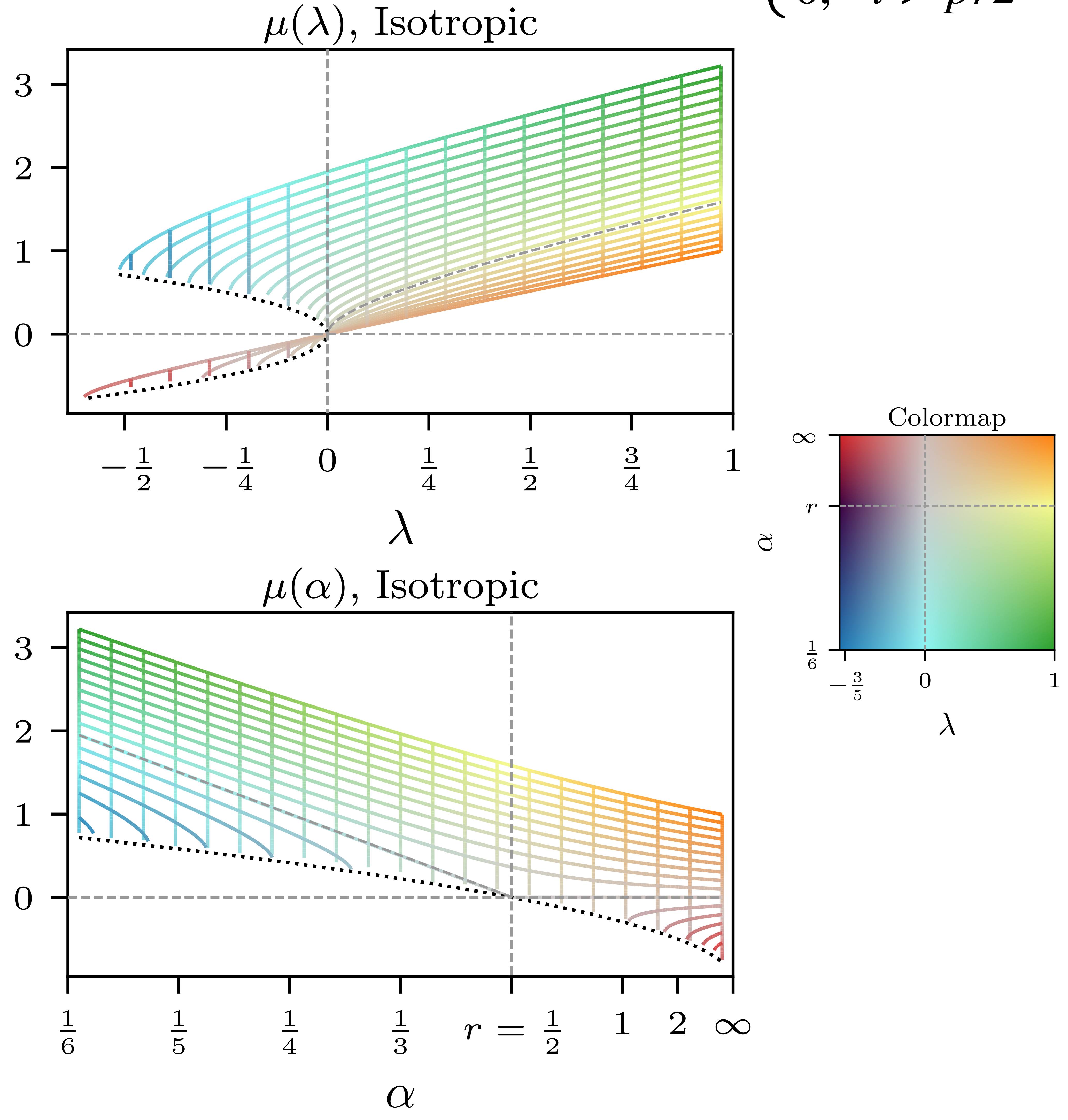


Some intuition about $\lambda \mapsto \mu$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$$

- Concave and increasing in λ

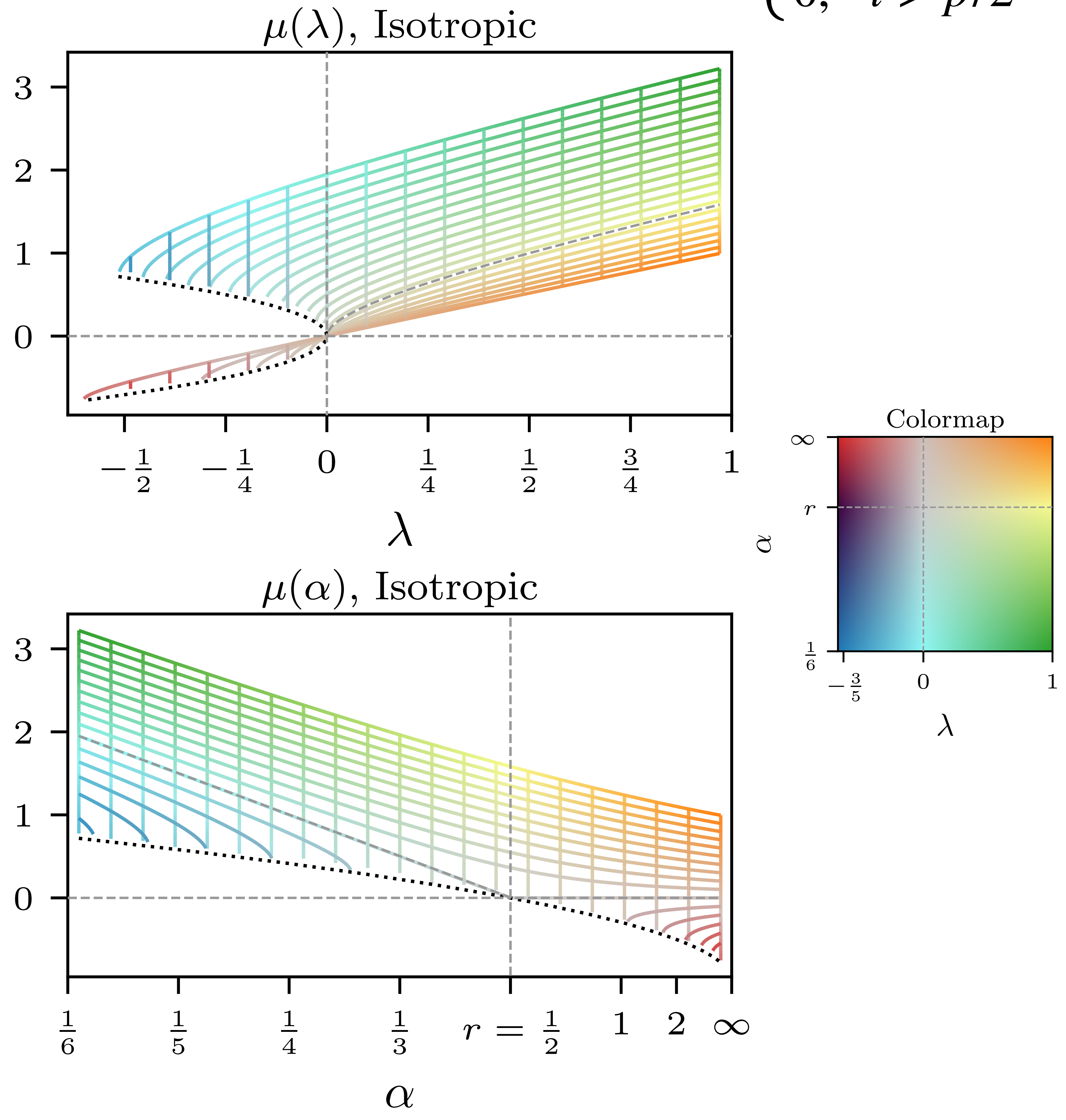
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$



Some intuition about $\lambda \mapsto \mu$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \le p/2 \\ 0, & i > p/2 \end{cases}$$

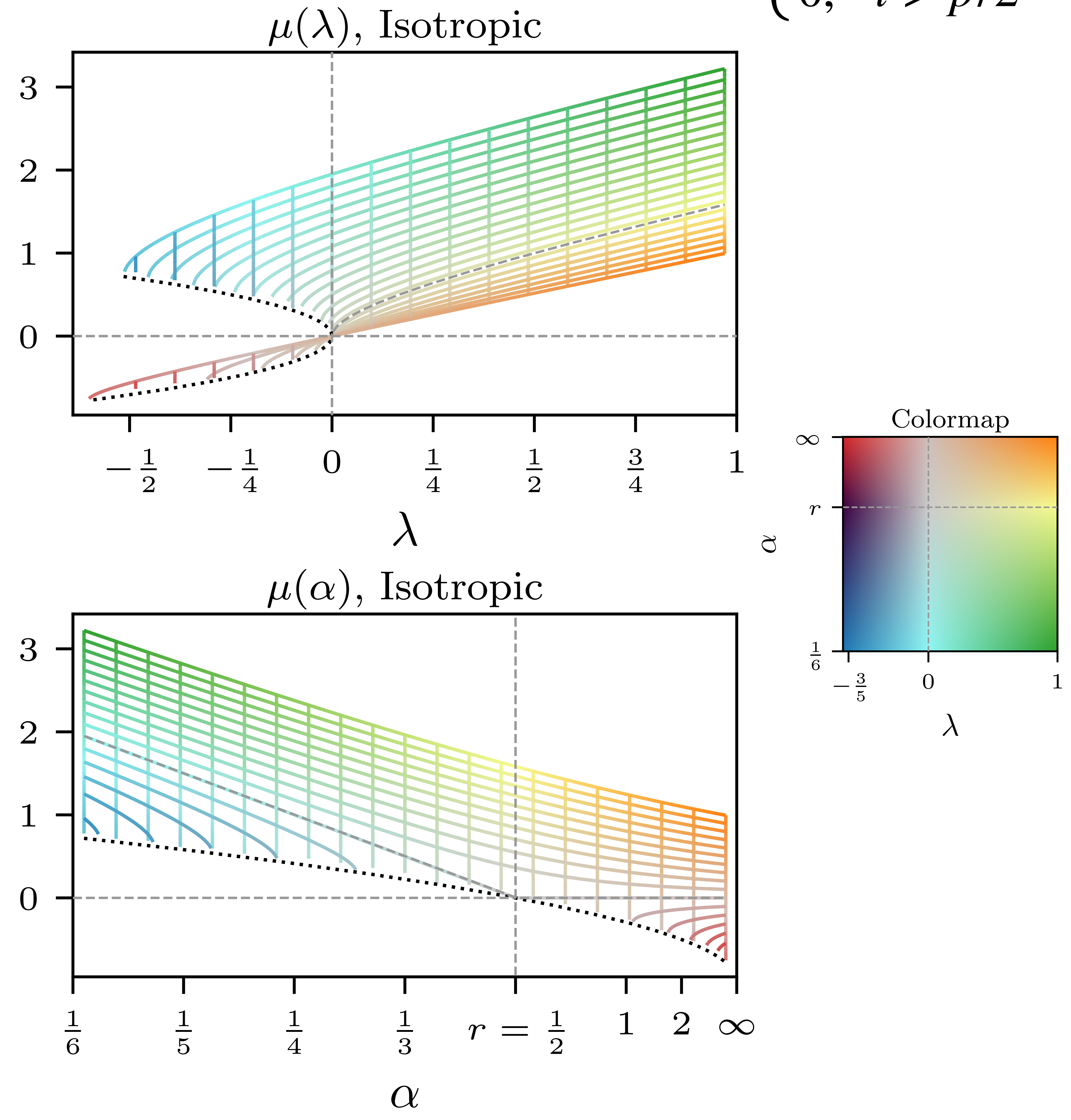
- Concave and increasing in λ
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $|\mu|$ decreasing in α for fixed λ



Some intuition about $\lambda \mapsto \mu$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$$

- Concave and increasing in λ
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $|\mu|$ decreasing in α for fixed λ
- $\lambda > 0$ or $q < \mathrm{rank}(\mathbf{A}) \implies \mu \geq 0, \mu > \lambda$



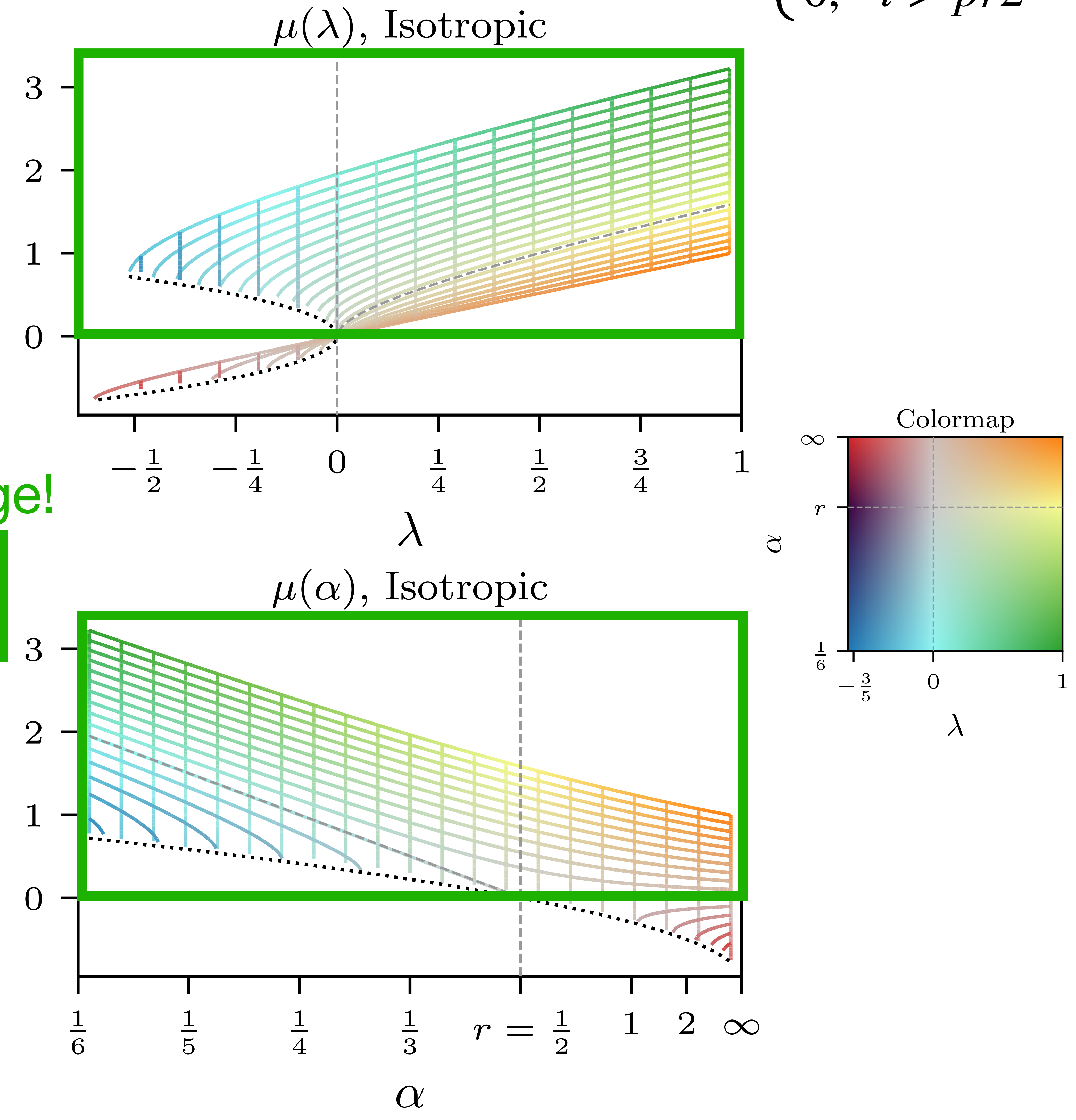
Some intuition about $\lambda \mapsto \mu$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \le p/2 \\ 0, & i > p/2 \end{cases}$$

- Concave and increasing in λ
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $|\mu|$ decreasing in α for fixed λ

sketching always adds ridge!

- $\lambda > 0$ or $q < \mathrm{rank}(\mathbf{A}) \implies \mu \geq 0, \mu > \lambda$



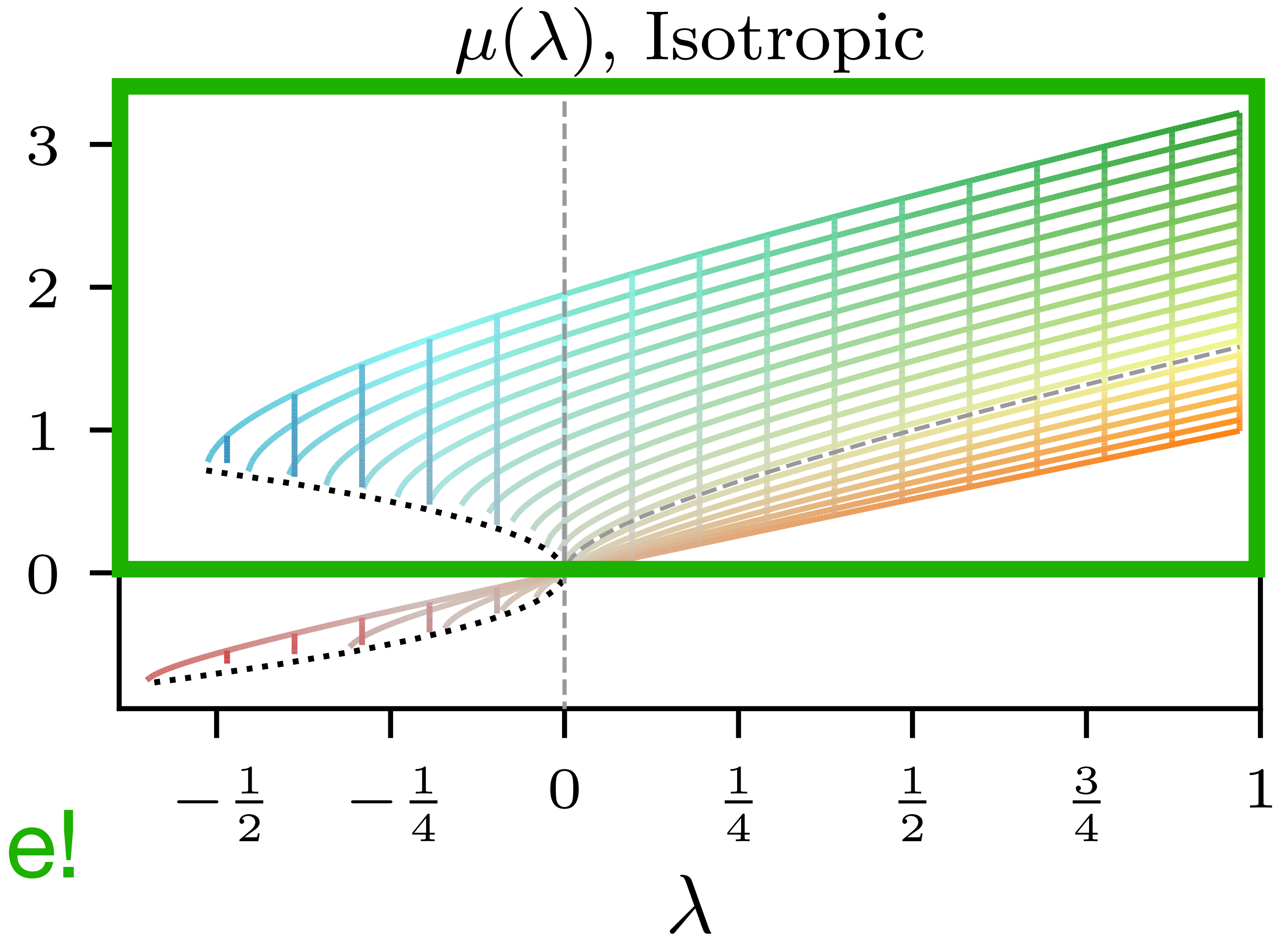
Some intuition about $\lambda \mapsto \mu$

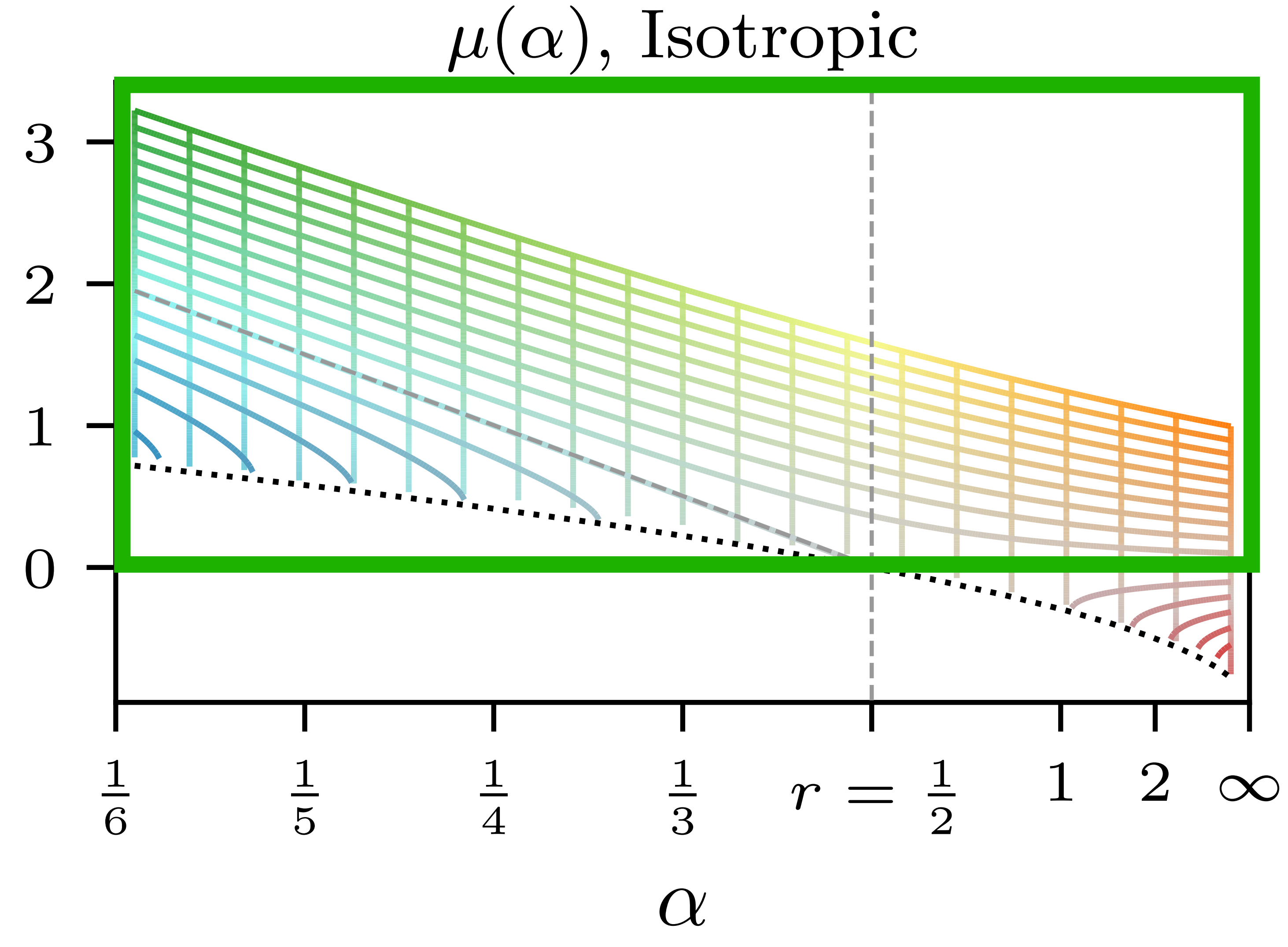
$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \le p/2 \\ 0, & i > p/2 \end{cases}$$

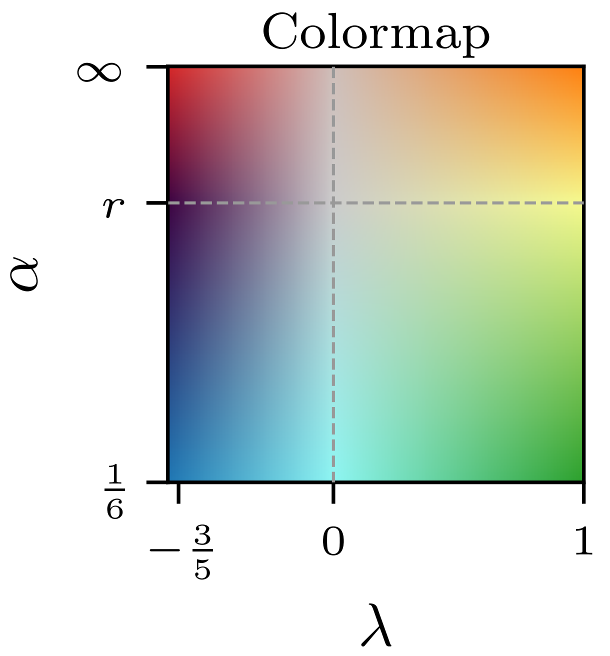
- Concave and increasing in λ
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $|\mu|$ decreasing in α for fixed λ

sketching always adds ridge!

- $\lambda > 0$ or $q < \mathrm{rank}(\mathbf{A}) \implies \mu \geq 0, \mu > \lambda$
- $\lambda < 0$ and $q > \mathrm{rank}(\mathbf{A}) \implies \mu < 0$







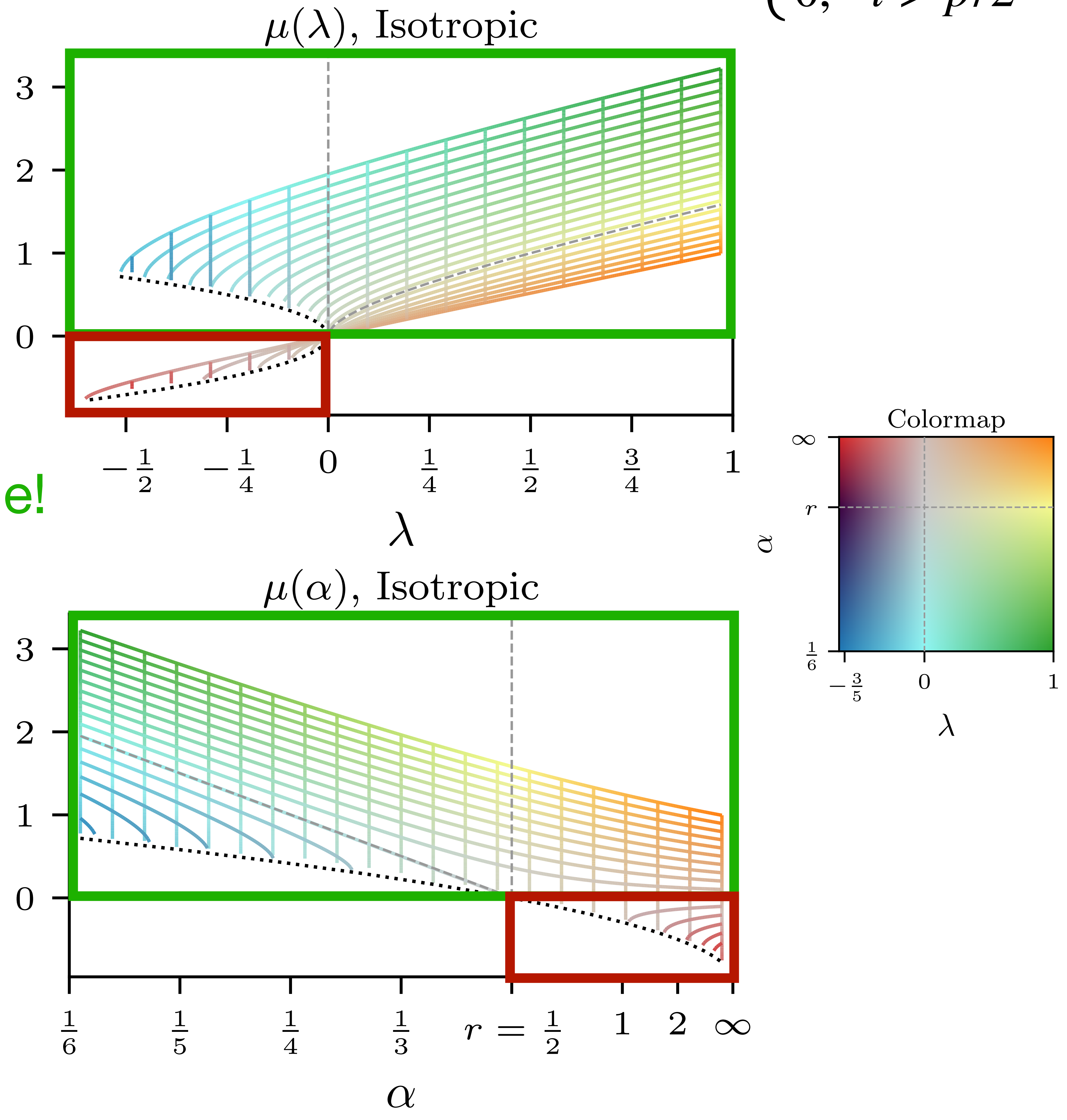
Some intuition about $\lambda \mapsto \mu$

- Concave and increasing in λ
- Limiting behavior: $\mu \sim \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $|\mu|$ decreasing in α for fixed λ

sketching always adds ridge!

- $\lambda > 0$ or $q < \mathrm{rank}(\mathbf{A}) \implies \mu \geq 0, \mu > \lambda$
- $\lambda < 0$ and $q > \mathrm{rank}(\mathbf{A}) \implies \mu < 0$

$$\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$$



Ridge regression risk?

Ridge regression risk?

- **Bias:** $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}^\top\mathbf{y} \simeq (\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I}_q)^{-1}\mathbf{X}^\top\mathbf{y} = \widehat{\mathbf{b}}_\mu$

Ridge regression risk?

- **Bias:** $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}^\top\mathbf{y} \simeq (\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I}_q)^{-1}\mathbf{X}^\top\mathbf{y} = \widehat{\mathbf{b}}_\mu$

- What about risk?

Ridge regression risk?

- **Bias:** $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}^\top\mathbf{y} \simeq (\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I}_q)^{-1}\mathbf{X}^\top\mathbf{y} = \widehat{\mathbf{b}}_\mu$
- What about risk?
  - **Problem:** $\widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu \not\Longrightarrow \widehat{\mathbf{b}}_{\mathbf{S}}^\top\widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu^\top\widehat{\mathbf{b}}_\mu$

Ridge regression risk?

- **Bias:** $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top \mathbf{X}^\top \mathbf{y} \simeq (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I}_q)^{-1} \mathbf{X}^\top \mathbf{y} = \widehat{\mathbf{b}}_\mu$
- What about risk?
  - **Problem:** $\widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu \;\not\Longrightarrow\; \widehat{\mathbf{b}}_{\mathbf{S}}^\top \widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$
  - Just like $\mathbb{E}[X] = \mathbb{E}[Y] \;\not\Longrightarrow\; \mathbb{E}[X^2] = \mathbb{E}[Y^2]$

Ridge regression risk?

- **Bias:** $\widehat{\mathbf{b}}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^\top \mathbf{X}^\top \mathbf{X}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \mathbf{X}^\top \mathbf{y} \simeq (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I}_q)^{-1}\mathbf{X}^\top \mathbf{y} = \widehat{\mathbf{b}}_\mu$
- What about risk?
  - **Problem:** $\widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu \quad \not\Longrightarrow \quad \widehat{\mathbf{b}}_{\mathbf{S}}^\top \widehat{\mathbf{b}}_{\mathbf{S}} \simeq \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$
  - Just like $\mathbb{E}[X] = \mathbb{E}[Y] \quad \not\Longrightarrow \quad \mathbb{E}[X^2] = \mathbb{E}[Y^2]$
  - We need a **second order** equivalence to work out **variance**

A second-order asymptotic equivalence

A second-order asymptotic equivalence

- **Theorem** (LeJeune, **PP**, et al., 2024). For any $\mathbf{\Psi}$ with uniformly bounded operator norm independent of $\mathbf{S}$ and the previous conditions, for i.i.d. $\mathbf{S}$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \mathbf{\Psi}\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1}(\mathbf{\Psi} + \mu' \mathbf{I}_p)(\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

where

$$\mu' = \frac{\frac{1}{q}\operatorname{tr}[\mu^3 \mathbf{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]}{\lambda + \frac{1}{q}\operatorname{tr}[\mu^2 \mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \geq 0.$$

A second-order asymptotic equivalence

- **Theorem** (LeJeune, **PP**, et al., 2024). For any $\boldsymbol{\Psi}$ with uniformly bounded operator norm independent of $\mathbf{S}$ and the previous conditions, for i.i.d. $\mathbf{S}$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \boldsymbol{\Psi} \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1}(\boldsymbol{\Psi} + \boxed{\mu' \mathbf{I}_p})(\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

sketching adds variance

where

$$\mu' = \frac{\frac{1}{q}\mathrm{tr}[\mu^3 \boldsymbol{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]}{\lambda + \frac{1}{q}\mathrm{tr}[\mu^2 \mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \geq 0.$$

A second-order asymptotic equivalence

- **Theorem** (LeJeune, **PP**, et al., 2024). For any $\boldsymbol{\Psi}$ with uniformly bounded operator norm independent of $\mathbf{S}$ and the previous conditions, for i.i.d. $\mathbf{S}$,

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \boldsymbol{\Psi} \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1}\mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1}(\boldsymbol{\Psi} + \mu' \mathbf{I}_p)(\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

sketching adds variance

where

$$\mu' = \frac{\frac{1}{q}\mathrm{tr}[\mu^3 \boldsymbol{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]}{\lambda + \frac{1}{q}\mathrm{tr}[\mu^2 \mathbf{A}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \geq 0.$$

- **Proof idea:** $\frac{\partial}{\partial z}(\mathbf{A} - z\mathbf{I})^{-1} = (\mathbf{A} - z\mathbf{I})^{-2}$ with carefully placed $\boldsymbol{\Psi}$

Ridge regression risk

Ridge regression risk

- Consider quadratic functional $R(\widehat{\mathbf{b}}_{\mathbf{S}}) = \widehat{\mathbf{b}}_{\mathbf{S}}^\top \boldsymbol{\Psi} \widehat{\mathbf{b}}_{\mathbf{S}} + \mathbf{h}^\top \widehat{\mathbf{b}}_{\mathbf{S}} + c$

Ridge regression risk

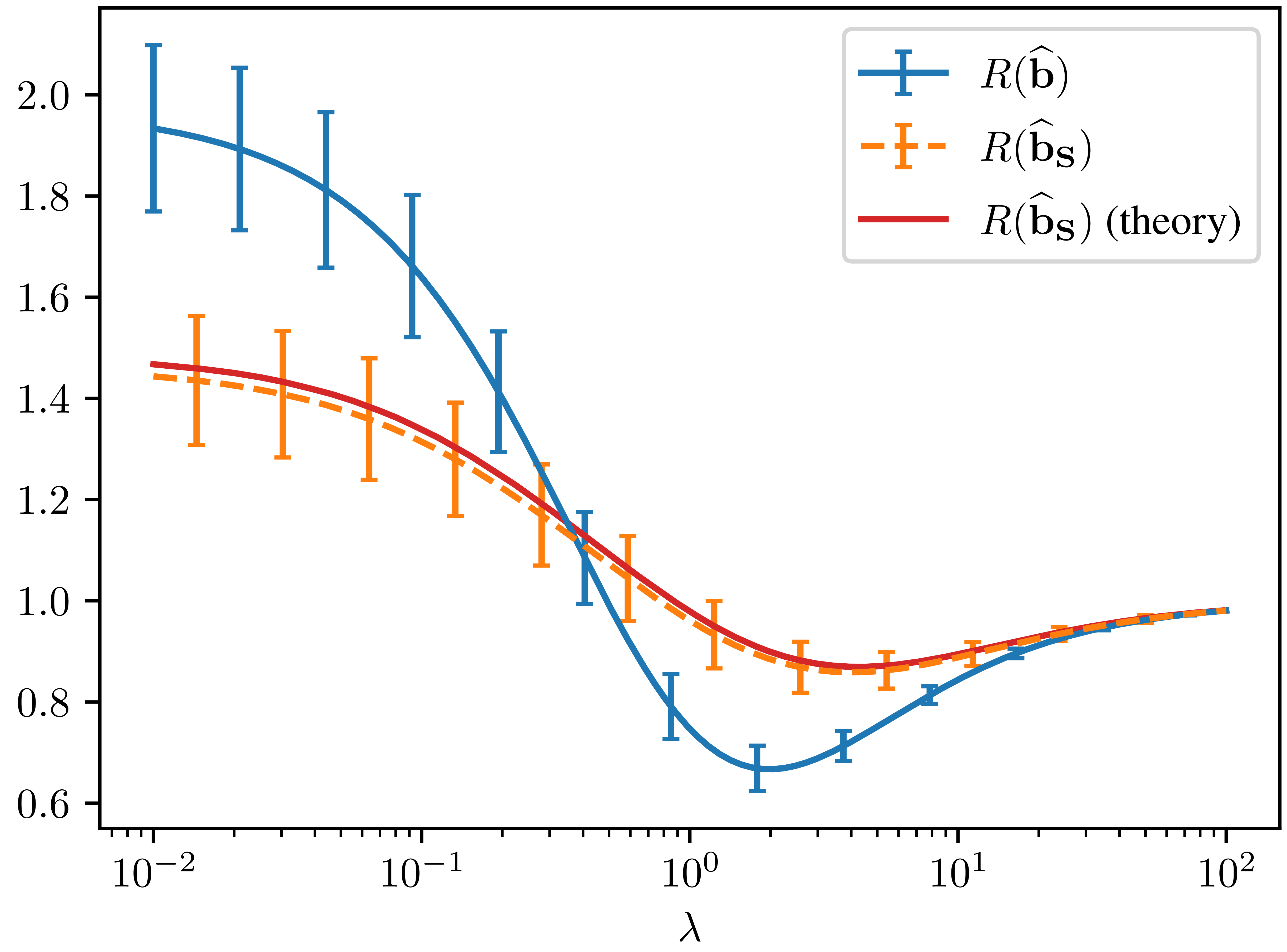
- Consider quadratic functional $R(\widehat{\mathbf{b}}_{\mathbf{S}}) = \widehat{\mathbf{b}}_{\mathbf{S}}^{\top} \boldsymbol{\Psi} \widehat{\mathbf{b}}_{\mathbf{S}} + \mathbf{h}^{\top} \widehat{\mathbf{b}}_{\mathbf{S}} + c$
  - Recall $\widehat{\mathbf{b}}_{\mu} = (\mathbf{X}^{\top}\mathbf{X} + \mu \mathbf{I}_p)^{-1} \mathbf{X}^{\top} \mathbf{y}$

Ridge regression risk

- Consider quadratic functional $R(\widehat{\mathbf{b}}_{\mathbf{S}}) = \widehat{\mathbf{b}}_{\mathbf{S}}^\top \boldsymbol{\Psi} \widehat{\mathbf{b}}_{\mathbf{S}} + \mathbf{h}^\top \widehat{\mathbf{b}}_{\mathbf{S}} + c$
  - Recall $\widehat{\mathbf{b}}_\mu = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$
  - Then $R(\widehat{\mathbf{b}}_{\mathbf{S}}) \simeq R(\widehat{\mathbf{b}}_\mu) + \mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$

Ridge regression risk

- Consider quadratic functional $R(\widehat{\mathbf{b}}_{\mathbf{S}}) = \widehat{\mathbf{b}}_{\mathbf{S}}^\top \boldsymbol{\Psi} \widehat{\mathbf{b}}_{\mathbf{S}} + \mathbf{h}^\top \widehat{\mathbf{b}}_{\mathbf{S}} + c$
  - Recall $\widehat{\mathbf{b}}_\mu = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$
  - Then $R(\widehat{\mathbf{b}}_{\mathbf{S}}) \simeq R(\widehat{\mathbf{b}}_\mu) + \mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$



When is sketching good/useful?

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\mathrm{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \frac{1}{q}\mathrm{tr}[\mu^2 \boldsymbol{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]\frac{\partial \mu}{\partial \lambda}$

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \frac{1}{q}\text{tr}[\mu^2 \mathbf{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}] \frac{\partial \mu}{\partial \lambda}$
    - α should be large

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \frac{1}{q}\text{tr}[\mu^2 \mathbf{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]\frac{\partial \mu}{\partial \lambda}$
    - α should be large
    - not much control via λ

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \frac{1}{q} \boxed{\mathrm{tr}[\mu^2 \mathbf{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \frac{\partial \mu}{\partial \lambda}$

    $\geq C$ if $\mathrm{rank}(\mathbf{\Psi}) > \mathrm{rank}(\mathbf{A})$

    - α should be large
    - not much control via λ

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \dfrac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \dfrac{1}{q} \mathrm{tr}[\mu^2 \boldsymbol{\Psi} (\mathbf{A} + \mu \mathbf{I}_p)^{-2}] \dfrac{\partial \mu}{\partial \lambda}$
    - α should be large
    - not much control via λ

$\geq C$ if $\mathrm{rank}(\boldsymbol{\Psi}) > \mathrm{rank}(\mathbf{A})$

≥ 1

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \dfrac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \dfrac{1}{q} \boxed{\text{tr}[\mu^2 \mathbf{\Psi} (\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \boxed{\dfrac{\partial \mu}{\partial \lambda}}$
    - $\geq C$ if $\text{rank}(\mathbf{\Psi}) > \text{rank}(\mathbf{A})$
    - ≥ 1
    - α should be large
    - not much control via λ
    - unless $\mu = 0$ and $\text{range}(\mathbf{\Psi}) \subseteq \text{range}(\mathbf{A})$!

When is sketching good/useful?

- When both $R(\widehat{\mathbf{b}}_\mu)$ and $\mu' \widehat{\mathbf{b}}_\mu^\top \widehat{\mathbf{b}}_\mu$ are small
  - $R(\widehat{\mathbf{b}}_\mu)$: λ should be less than $\mu = \lambda_{\text{opt}}$, while essentially $\alpha \propto \frac{1}{\mu - \lambda}$
  - μ': consider alternate form $\mu' = \frac{1}{q} \boxed{\operatorname{tr}[\mu^2 \boldsymbol{\Psi}(\mathbf{A} + \mu \mathbf{I}_p)^{-2}]} \boxed{\frac{\partial \mu}{\partial \lambda}}$

    $\geq C$ if $\operatorname{rank}(\boldsymbol{\Psi}) > \operatorname{rank}(\mathbf{A})$

    ≥ 1
    - α should be large
    - not much control via λ
    - unless $\mu = 0$ and $\operatorname{range}(\boldsymbol{\Psi}) \subseteq \operatorname{range}(\mathbf{A})$!
      - requires $\lambda = 0$ and $q > \operatorname{rank}(\mathbf{A})$

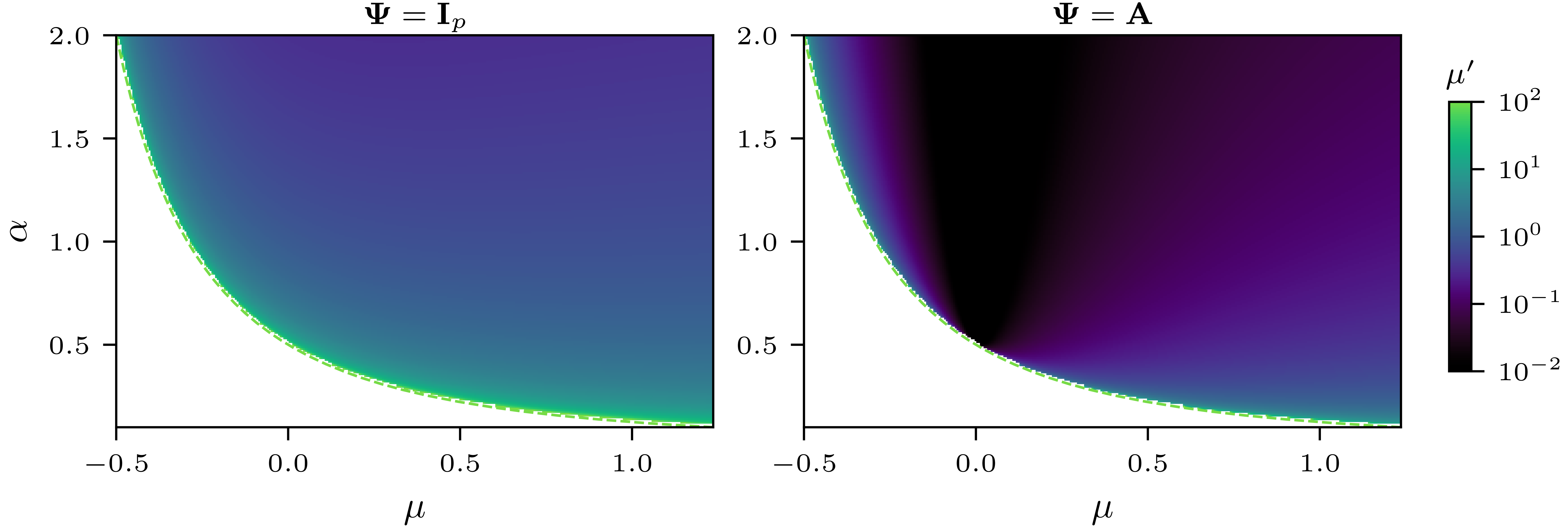
The magic of sketched ridgeless regression

The magic of sketched ridgeless regression

- Example: rank-deficient isotropy, $\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$

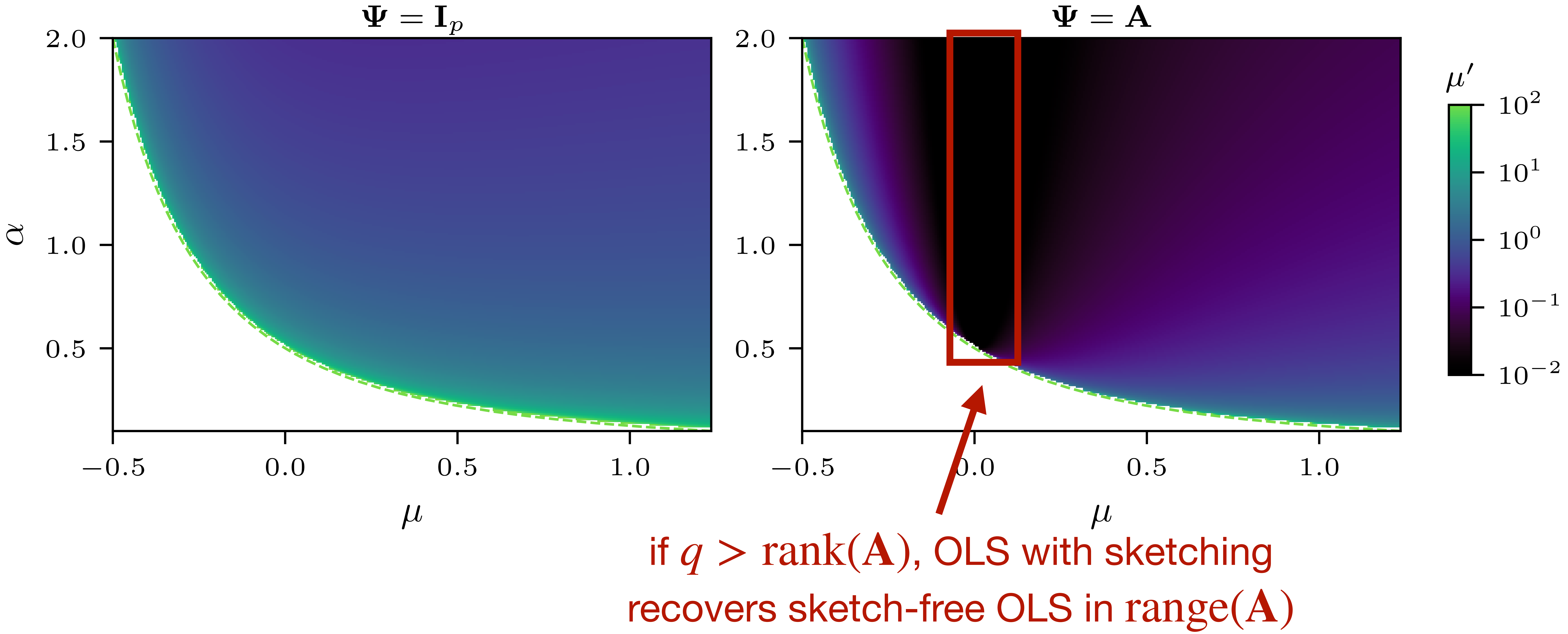
The magic of sketched ridgeless regression

- Example: rank-deficient isotropy, $\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$



The magic of sketched ridgeless regression

- Example: rank-deficient isotropy, $\lambda_i(\mathbf{A}) = \begin{cases} 1, & i \leq p/2 \\ 0, & i > p/2 \end{cases}$



Consistent risk estimation

Consistent risk estimation

- **Generalized cross-validation (GCV)**

Consistent risk estimation

- **Generalized cross-validation (GCV)**
- $$\hat{R}(\widehat{\mathbf{b}}_{\mathbf{S}}) = \frac{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}_{\mathbf{S}}\|_2^2}{(1 - \frac{1}{n}\mathrm{tr}[\mathbf{X}\mathbf{S}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}]^\top)^2}$$

Consistent risk estimation

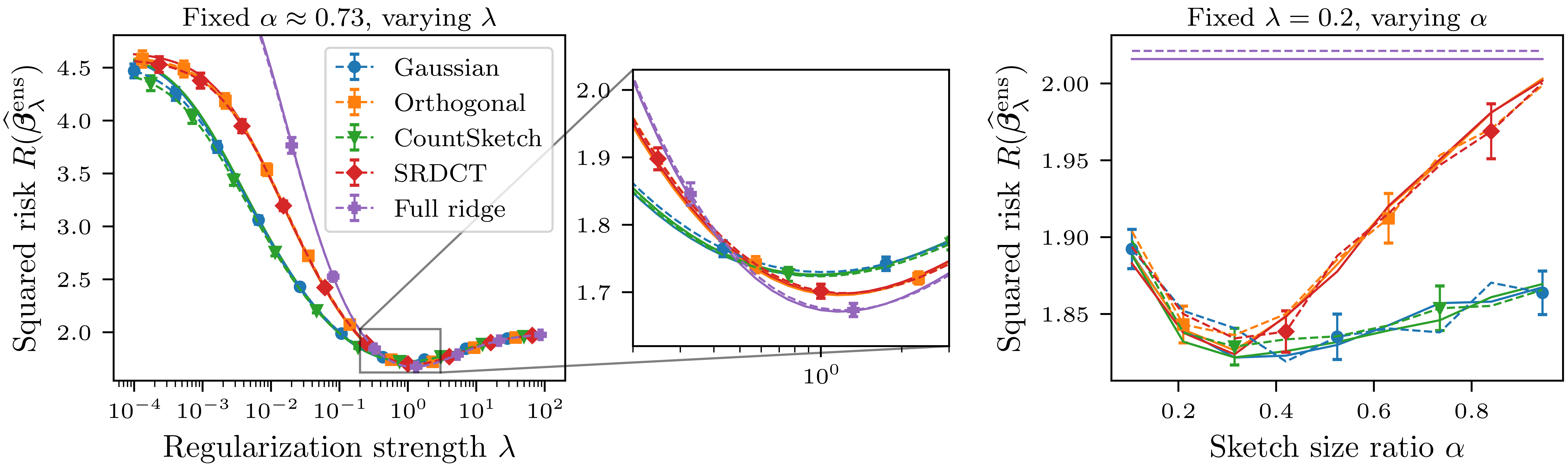
- **Generalized cross-validation (GCV)**
  - $$\hat{R}(\widehat{\mathbf{b}}_{\mathbf{S}}) = \frac{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}_{\mathbf{S}}\|_2^2}{(1 - \frac{1}{n}\mathrm{tr}[\mathbf{X}\mathbf{S}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}]^\top)^2}$$
  - Costs the same as $\widehat{\mathbf{b}}_{\mathbf{S}}$ to compute

Consistent risk estimation

- **Generalized cross-validation (GCV)**
  - $$\hat{R}(\widehat{\mathbf{b}}_{\mathbf{S}}) = \frac{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}_{\mathbf{S}}\|_2^2}{(1 - \frac{1}{n}\mathrm{tr}[\mathbf{XS}(\mathbf{S}^\top\mathbf{X}^\top\mathbf{XS} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^\top\mathbf{X}]^\top)^2}$$
  - Costs the same as $\widehat{\mathbf{b}}_{\mathbf{S}}$ to compute
- **Theorem** (**PP** & LeJeune, 2024). For any asymptotically free sketch $\mathbf{S}$, under random data assumptions on $\mathbf{X}$,

$$\hat{R}(\widehat{\mathbf{b}}_{\mathbf{S}}) \simeq R(\widehat{\mathbf{b}}_{\mathbf{S}}) \simeq R(\widehat{\mathbf{b}}_{\mu}) + \mu'\Delta.$$

Consistent risk estimation



Ensemble trick for unsketched risk

Ensemble trick for unsketched risk

- Let $\widehat{\mathbf{b}}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X}\mathbf{S}_k + \lambda \mathbf{I}_q)^{-1}\mathbf{X}^\top \mathbf{y}$

Ensemble trick for unsketched risk

- Let $\widehat{\mathbf{b}}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X}\mathbf{S}_k + \lambda \mathbf{I}_q)^{-1}\mathbf{X}^\top \mathbf{y}$

- Then $\hat{R}(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_\mu) + \frac{\mu'}{K}\Delta$

Ensemble trick for unsketched risk

- Let $\widehat{\mathbf{b}}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X}\mathbf{S}_k + \lambda \mathbf{I}_q)^{-1}\mathbf{X}^\top \mathbf{y}$
- Then $\hat{R}(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_\mu) + \frac{\mu'}{K}\Delta$
- Given the mapping $\lambda \mapsto \mu$, this admits a consistent estimator

$$R(\widehat{\mathbf{b}}_\mu) \simeq 2\hat{R}(\widehat{\mathbf{b}}_{K=2}) - \hat{R}(\widehat{\mathbf{b}}_{K=1})$$

Ensemble trick for unsketched risk

- Let $\widehat{\mathbf{b}}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X}\mathbf{S}_k + \lambda \mathbf{I}_q)^{-1}\mathbf{X}^\top \mathbf{y}$

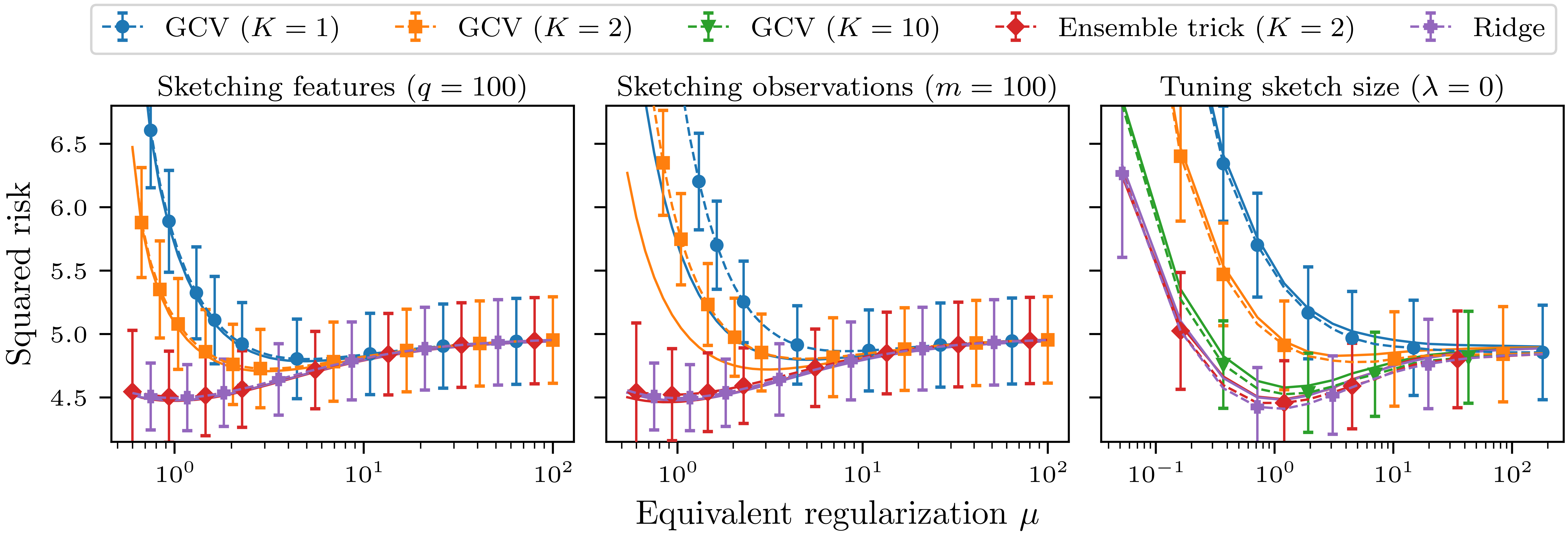
- Then $\hat{R}(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_K) \simeq R(\widehat{\mathbf{b}}_\mu) + \frac{\mu'}{K}\Delta$

- Given the mapping $\lambda \mapsto \mu$, this admits a consistent estimator

$$R(\widehat{\mathbf{b}}_\mu) \simeq 2\hat{R}(\widehat{\mathbf{b}}_{K=2}) - \hat{R}(\widehat{\mathbf{b}}_{K=1})$$

- Cost (for iterative solver) is $\mathcal{O}(4nq)$ versus $\mathcal{O}(2np)$, efficient if $q \ll p$

Ensemble trick for unsketched risk



Summary

Summary

- **Our contribution:**

Summary

- **Our contribution:**
    - Precise characterization of implicit regularization μ of sketching

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk
  - Consistency of GCV for sketched ridge regression

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk
  - Consistency of GCV for sketched ridge regression
    - Efficient risk estimation via ensemble trick

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk
  - Consistency of GCV for sketched ridge regression
    - Efficient risk estimation via ensemble trick
- **Ongoing work:**

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk
  - Consistency of GCV for sketched ridge regression
    - Efficient risk estimation via ensemble trick
- **Ongoing work:**
  - Applying first & second order analysis to sketch-and-project

Summary

- **Our contribution:**
  - Precise characterization of implicit regularization μ of sketching
    - Covers free sketches $\mathbf{S}$, any data $\mathbf{A}$, sketch ratio α, (negative) λ
    - Includes second order characterization for risk
  - Consistency of GCV for sketched ridge regression
    - Efficient risk estimation via ensemble trick
- **Ongoing work:**
  - Applying first & second order analysis to sketch-and-project
  - Efficient risk estimation for general learning problems

- Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space.
- Lu, Y., Dhillon, P., Foster, D. P., and Ungar, L. (2013). Faster ridge regression via the subsampled randomized Hadamard transform.
- Thanei, G. A., Heinze, C., and Meinshausen, N. (2017). Random projections for large-scale regression.
- Dobriban, E., and Sheng, Y. (2021). Distributed linear regression by averaging.
- LeJeune, D., **PP**, Javadi, H., Baraniuk, R. G., and Tibshirani, R. J. (2024). Asymptotics of the sketched pseudoinverse.
- **PP**, LeJeune, D. (2024). Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning.