1 Pseudo-linear form

Derivation of Peyman Milanfar's gradient

$$
\begin{aligned}
\mathrm{d}[\mathbf{f}(\mathbf{x})] &= \mathrm{d}[\mathbf{A}(\mathbf{x})\mathbf{x}] \\
&= \mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x} + \mathbf{A}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \mathrm{vec}\{\mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \mathrm{vec}\{\mathbf{I}\mathrm{d}[\mathbf{A}(\mathbf{x})]\mathbf{x}\} + \mathbf{A}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \left(\mathbf{x}^T \otimes \mathbf{I}\right) \mathrm{vec}\{\mathrm{d}[\mathbf{A}(\mathbf{x})]\} + \mathbf{A}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \left(\mathbf{x}^T \otimes \mathbf{I}\right) \mathrm{D}\,\mathrm{vec}[\mathbf{A}(\mathbf{x})]\mathrm{d}\mathbf{x} + \mathbf{A}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \left[\left(\mathbf{x}^T \otimes \mathbf{I}\right) \mathrm{D}\,\mathrm{vec}[\mathbf{A}(\mathbf{x})] + \mathbf{A}(\mathbf{x})\right] \mathrm{d}\mathbf{x}
\end{aligned}
$$

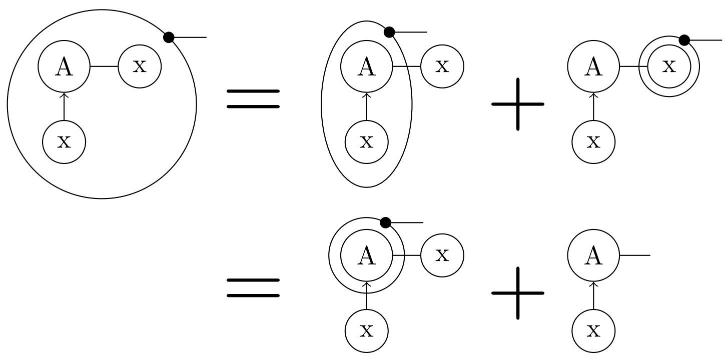

Figure 1: Visualization of pseudo-linear gradient.

Note, a third way to derive the gradient is to use index notation:

$$
\begin{aligned}
f_i(\mathbf{x}) &= A_{ij}(\mathbf{x})x_j \\
\Rightarrow \mathrm{d}f_i &= \frac{\partial f_i}{\partial x_k}\mathrm{d}x_k \\
&= \left(\frac{\partial A_{ij}}{\partial x_k}x_j + A_{ij}\delta_{jk}\right)\mathrm{d}x_k \\
&= \left(\frac{\partial A_{ij}}{\partial x_k}x_j + A_{ik}\right)\mathrm{d}x_k
\end{aligned}
$$

2 Chain Rule

Standard chain rule. Here we let $f \in \mathbb{R}^d \to \mathbb{R}$ be a scalar function, and $v \in \mathbb{R}^d \to \mathbb{R}^d$ be a vector function as used in backprop.

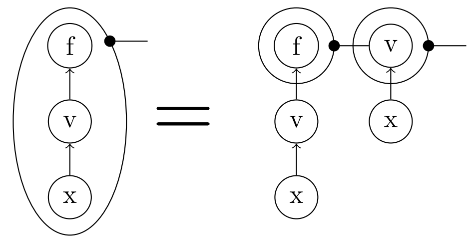

Figure 2: Visualization of the Chain Rule: $J_{f\circ v}(x) = \nabla_f(v(x))J_v(x)$.

3 Computation of the Hessian

Derivation of Yaroslav Bulatov's chain rule for the Hessian. See Figure 3.

In index notation, the Hessian of $f(v(x))$ is

$$H_{ij}(x) = \sum_{k=1}^{d}\sum_{l=1}^{d} \frac{\partial^2 f}{\partial u_k \partial u_l}(v(x))\frac{\partial v_k}{\partial x_i}(x)\frac{\partial v_l}{\partial x_j}(x) + \frac{\partial f}{\partial u_k}(v(x))\frac{\partial^2 v_k}{\partial x_i \partial x_j}(x).$$

In matrix notation it is

$$H(x) = Dv(x)^T \cdot D^2 f(v(x)) \cdot Dv(x) + \sum_{k=1}^{d} \frac{\partial f}{\partial u_k}(v(x))\frac{\partial^2 v_k}{\partial x \partial x^T}(x).$$

Neither of them are terribly legible.

4 Quadratic form

A common gradient from statistics, is the least squares $\nabla_x \|Ax-b\|_2^2 = \nabla_x (Ax)^T(Ax) - 2b^T Ax + b^T b$. See Figure 4.

Once the gradient has been derived, we can solve for x to get the usual solution $x = (A^T A)^{-1}Ab$.

5 Quadratic form 2

In machine learning we sometimes want a "matrix shaped" gradient that we can easily add to the original matrix for gradient descent. Let's define a derivative notation with two edges going out for this purpose. Then we can derive the gradient with respect to X of $\nabla_X \|Xa - b\|_2^2$. See Figure 5.

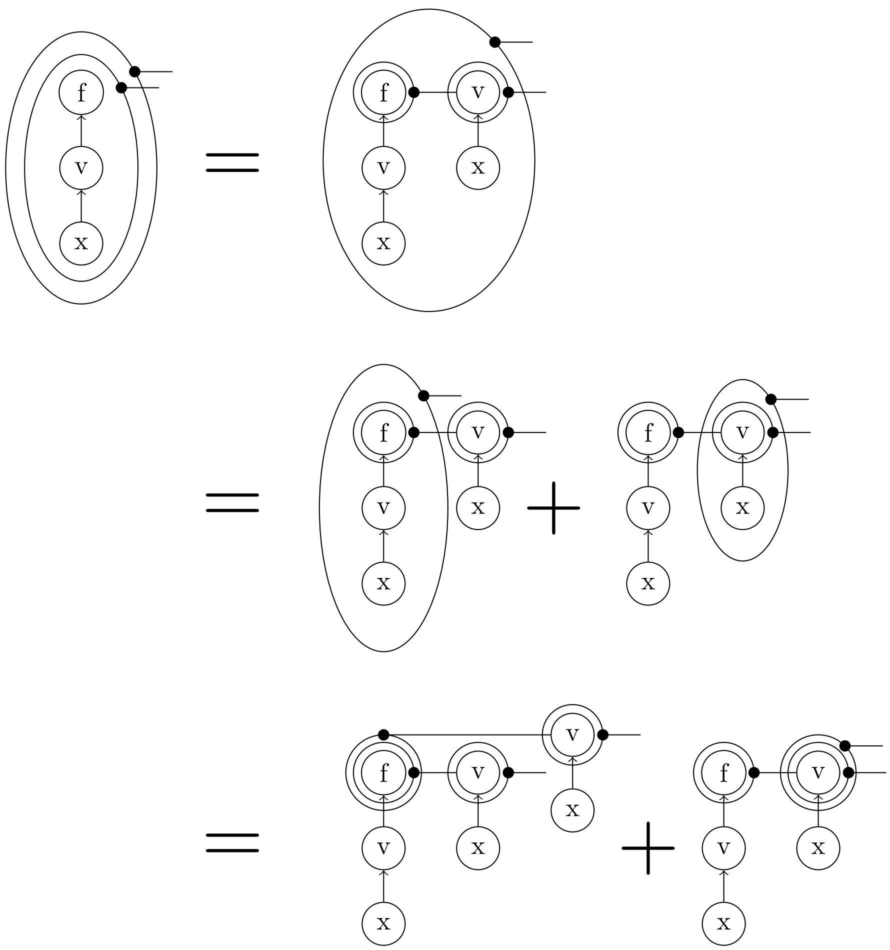

Figure 3: Visualization of the Computation of the Hessian: $H_{f \circ v}(x) = Dv(x)^T \cdot D^2 f(v(x)) \cdot Dv(x) + \sum_{k=1}^{d} \frac{\partial f}{\partial u_k}(v(x)) \frac{\partial^2 v_k}{\partial x \partial x^T}(x)$..

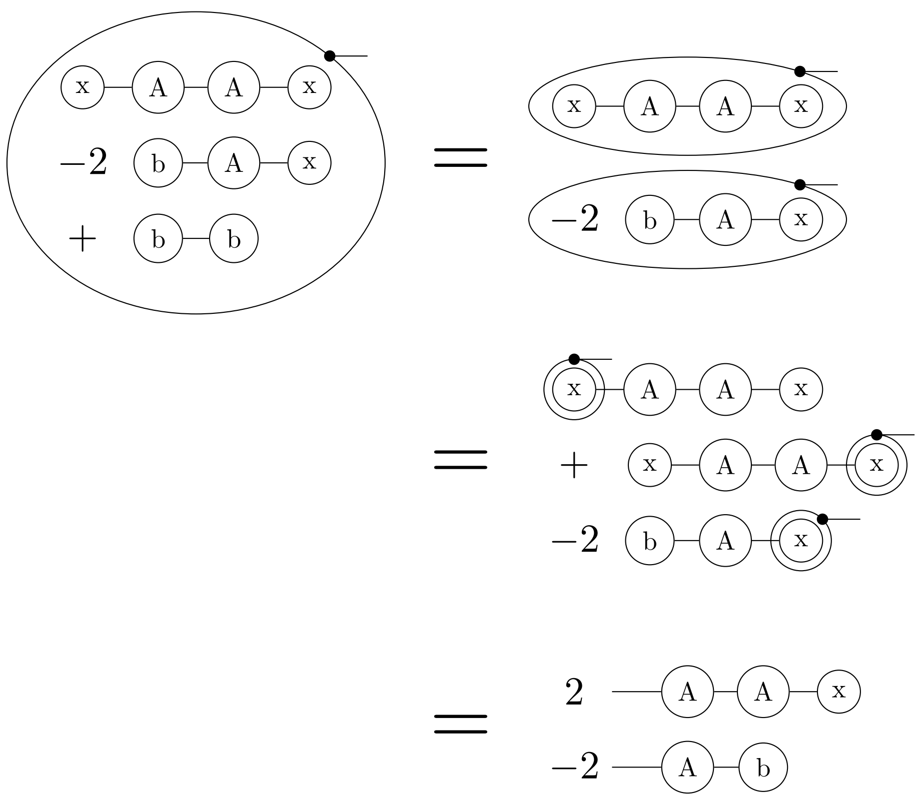

Figure 4: Least squares gradient, $\nabla_x \|Ax - b\|_2^2 = 2A^T Ax - 2Ab$.

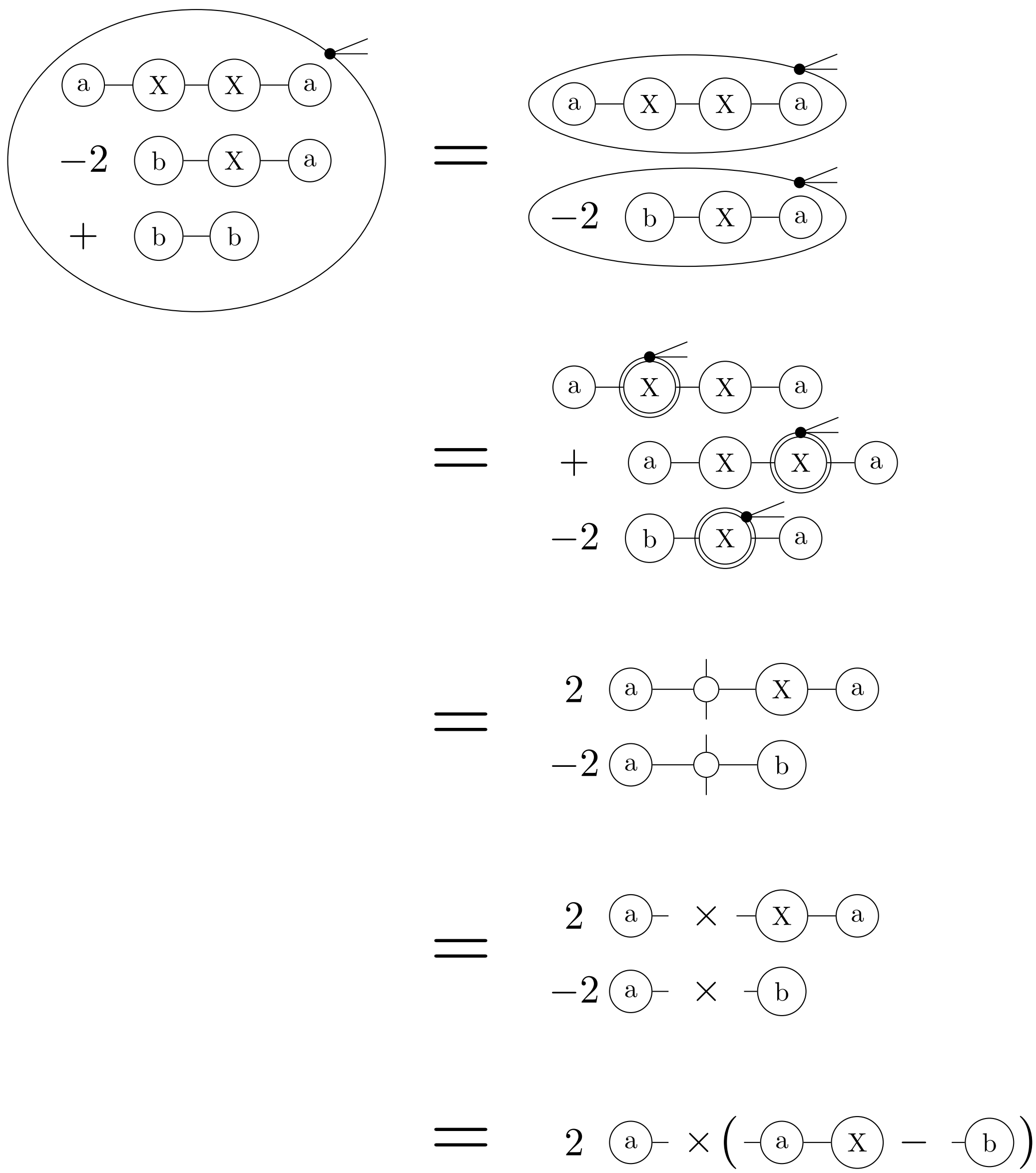

Figure 5: Least squares gradient, $\nabla_X \|aX - b\|_2^2 = 2a \otimes (Xa - b)$. In step four we used a small trick, which is that $x^T(I \otimes I)x = (I \otimes I)(x \otimes x) = (Ix) \otimes (Ix) = x \otimes x$. In other words, the degree 4 identity matrix splits into the outer product of it's constituents.